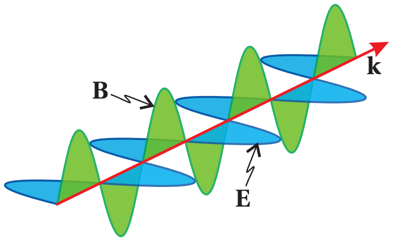
Physics of Light and Optics



Justin Peatross
Michael Ware
Brigham Young University

2015 Edition
June 25, 2021 Revision

Copyright ©2015 Justin Peatross and Michael Ware

All rights reserved. The authors retain the copyright to this book. However, the content is available free of charge at optics.byu.edu. This book may be downloaded, printed, and distributed freely as long this copyright notice is included. Any use of a portion of this book's content as part of another other work requires the express written permission of the authors.

ISBN 978-1-312-92927-2

Preface

This curriculum was originally developed for a fourth-year undergraduate optics course in the Department of Physics and Astronomy at Brigham Young University. Topics are addressed from a physics perspective and include the propagation of light in matter, reflection and transmission at boundaries, polarization effects, dispersion, coherence, ray optics and imaging, diffraction, and the quantum nature of light. Students using this book should be familiar with differentiation, integration, and standard trigonometric and algebraic manipulation. A brief review of complex numbers, vector calculus, and Fourier transforms is provided in Chapter 0, but it is helpful if students already have some experience with these concepts.

While the authors retain the copyright, we have made this book available free of charge at optics.byu.edu. This is our contribution toward a future world with free textbooks! The web site also provides a link to purchase bound copies of the book for the cost of printing. A collection of electronic material related to the text is available at the same site, including videos of students performing the lab assignments found in the book.

The development of optics has a rich history. We have included historical sketches for a selection of the pioneers in the field to help students appreciate some of this historical context. These sketches are not intended to be authoritative; the information for most individuals has been gleaned primarily from Wikipedia.

The authors may be contacted at opticsbook@byu.edu. We enjoy hearing reports from those using the book and welcome constructive feedback. We occasionally revise the text. The title page indicates the date of the last revision.

We would like to thank all those who have helped improve this material. We especially thank John Colton, Bret Hess, and Harold Stokes for their careful review and extensive suggestions. This curriculum benefitted from a CCLI grant from the National Science Foundation Division of Undergraduate Education (DUE-9952773).

Contents

Preface iii

Table of Contents v

0 Mathematical Tools 1
- 0.1 Vector Calculus 1
- 0.2 Complex Numbers 6
- 0.3 Linear Algebra 11
- 0.4 Fourier Theory 13
- Appendix 0.A Table of Integrals and Sums 19
- Exercises 20

1 Electromagnetic Phenomena 25
- 1.1 Gauss' Law 26
- 1.2 Gauss' Law for Magnetic Fields 27
- 1.3 Faraday's Law 29
- 1.4 Ampere's Law 30
- 1.5 Maxwell's Adjustment to Ampere's Law 31
- 1.6 Polarization of Materials 34
- 1.7 The Wave Equation 36
- Exercises 39

2 Plane Waves and Refractive Index 43
- 2.1 Plane Wave Solutions to the Wave Equation 43
- 2.2 Complex Plane Waves 45
- 2.3 Index of Refraction 46
- 2.4 The Lorentz Model of Dielectrics 49
- 2.5 Index of Refraction of a Conductor 52
- 2.6 Poynting's Theorem 54
- 2.7 Irradiance of a Plane Wave 56
- Appendix 2.A Radiometry, Photometry, and Color 58
- Appendix 2.B Clausius-Mossotti Relation 62
- Appendix 2.C Energy Density of Electric Fields 65
- Appendix 2.D Energy Density of Magnetic Fields 67
- Exercises 68

3 Reflection and Refraction **73**
3.1 Refraction at an Interface 73
3.2 The Fresnel Coefficients 77
3.3 Reflectance and Transmittance 78
3.4 Brewster's Angle 80
3.5 Total Internal Reflection 81
3.6 Reflections from Metal 83
Appendix 3.A Boundary Conditions For Fields at an Interface 84
Exercises 86

4 Multiple Parallel Interfaces **89**
4.1 Double-Interface Problem Solved Using Fresnel Coefficients . . . 90
4.2 Transmittance through Double Interface at Subcritical Angles . . 94
4.3 Beyond Critical Angle: Tunneling of Evanescent Waves 97
4.4 Fabry-Perot Instrument 98
4.5 Setup of a Fabry-Perot Instrument 100
4.6 Distinguishing Nearby Wavelengths in a Fabry-Perot Instrument . 102
4.7 Multilayer Coatings 105
4.8 Periodic Multilayer Stacks 109
Exercises 111

Review, Chapters 1–4 **115**

5 Propagation in Anisotropic Media **121**
5.1 Constitutive Relation in Crystals 121
5.2 Plane Wave Propagation in Crystals 123
5.3 Biaxial and Uniaxial Crystals 127
5.4 Refraction at a Uniaxial Crystal Surface 129
5.5 Poynting Vector in a Uniaxial Crystal 130
Appendix 5.A Symmetry of Susceptibility Tensor 132
Appendix 5.B Rotation of Coordinates 133
Appendix 5.C Electric Field in a Crystal 135
Appendix 5.D Huygens' Elliptical Construct for a Uniaxial Crystal . . . 138
Exercises 140

6 Polarization of Light **143**
6.1 Linear, Circular, and Elliptical Polarization 144
6.2 Jones Vectors for Representing Polarization 145
6.3 Elliptically Polarized Light 146
6.4 Linear Polarizers and Jones Matrices 147
6.5 Jones Matrix for a Polarizer 150
6.6 Jones Matrix for Wave Plates 151
6.7 Polarization Effects of Reflection and Transmission 153
Appendix 6.A Ellipsometry 155
Appendix 6.B Partially Polarized Light 156

Exercises 164

7 Superposition of Quasi-Parallel Plane Waves 169
7.1 Intensity of Superimposed Plane Waves 170
7.2 Group vs. Phase Velocity: Sum of Two Plane Waves 172
7.3 Frequency Spectrum of Light 174
7.4 Wave Packet Propagation and Group Delay 178
7.5 Quadratic Dispersion 181
7.6 Generalized Context for Group Delay 183
Appendix 7.A Pulse Chirping in a Grating Pair 187
Appendix 7.B Causality and Exchange of Energy with the Medium 189
Appendix 7.C Kramers-Kronig Relations 194
Exercises 198

8 Coherence Theory 203
8.1 Michelson Interferometer 203
8.2 Coherence Time and Fringe Visibility 207
8.3 Temporal Coherence of Continuous Sources 209
8.4 Fourier Spectroscopy 209
8.5 Young's Two-Slit Setup and Spatial Coherence 211
Appendix 8.A Spatial Coherence for a Continuous Spatial Distribution 215
Appendix 8.B Van Cittert-Zernike Theorem 216
Exercises 219

Review, Chapters 5–8 223

9 Light as Rays 229
9.1 The Eikonal Equation 230
9.2 Fermat's Principle 233
9.3 Paraxial Rays and ABCD Matrices 236
9.4 Reflection and Refraction at Curved Surfaces 238
9.5 ABCD Matrices for Combined Optical Elements 240
9.6 Image Formation 243
9.7 Principal Planes for Complex Optical Systems 246
9.8 Stability of Laser Cavities 248
Appendix 9.A Aberrations and Ray Tracing 250
Exercises 254

10 Diffraction 259
10.1 Huygens' Principle as Formulated by Fresnel 260
10.2 Scalar Diffraction Theory 262
10.3 Fresnel Approximation 264
10.4 Fraunhofer Approximation 266
10.5 Diffraction with Cylindrical Symmetry 267
Appendix 10.A Fresnel-Kirchhoff Diffraction Formula 269
Appendix 10.B Green's Theorem 272

Exercises 273

11 Diffraction Applications 277
11.1 Fraunhofer Diffraction with a Lens 277
11.2 Resolution of a Telescope 282
11.3 The Array Theorem 284
11.4 Diffraction Grating 286
11.5 Spectrometers 287
11.6 Diffraction of a Gaussian Field Profile 290
11.7 Gaussian Laser Beams 292
Appendix 11.A ABCD Law for Gaussian Beams 294
Exercises 298

12 Interferograms and Holography 305
12.1 Interferograms 305
12.2 Testing Optical Surfaces 306
12.3 Generating Holograms 307
12.4 Holographic Wavefront Reconstruction 308
Exercises 311

Review, Chapters 9–12 313

13 Blackbody Radiation 319
13.1 Stefan-Boltzmann Law 320
13.2 Failure of the Equipartition Principle 321
13.3 Planck's Formula 323
13.4 Einstein's A and B Coefficients 326
Appendix 13.A Thermodynamic Derivation of the Stefan-Boltzmann Law 328
Appendix 13.B Boltzmann Factor 330
Exercises 332

Index 335

Physical Constants 340

Chapter 0

Mathematical Tools

Before moving on to chapter 1 where our study of optics begins, it would be good to look over this chapter to make sure you are comfortable with the mathematical tools we'll be using. The vector calculus information in section 0.1 is used straight away in Chapter 1, so you should review it now. In Section 0.2 we review complex numbers. You have probably had some exposure to complex numbers, but if you are like many students, you haven't yet fully appreciated their usefulness. Your life will be *much easier* if you understand the material in section 0.2 by heart. Complex notation is pervasive throughout the book, beginning in chapter 2.

You may safely procrastinate reviewing Sections 0.3 and 0.4 until they come up in the book. The linear algebra refresher in Section 0.3 is useful for Chapter 4, where we analyze multilayer coatings, and again in Chapter 6, where we discuss polarization. Section 0.4 provides an introduction to Fourier theory. Fourier transforms are used extensively in optics, and you should study Section 0.4 carefully before tackling Chapter 7.

René Descartes (1596-1650, French) was born in La Haye en Touraine (now Descartes), France. His mother died when he was an infant. His father was a member of parliament who encouraged Descartes to become a lawyer. Descartes graduated with a degree in law from the University of Poitiers in 1616. In 1619, he had a series of dreams that led him to believe that he should instead pursue science. Descartes became one of the greatest mathematicians, physicists, and philosophers of all time. He is credited with inventing the Cartesian coordinate system, which is named after him. For the first time, geometric shapes could be expressed as algebraic equations. (Wikipedia)

0.1 Vector Calculus

Each position in space corresponds to a unique *vector* $\mathbf{r} \equiv x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}$, where $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are *unit vectors* with length one, pointing along their respective axes. Boldface type distinguishes a variable as a vector quantity, and the use of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ denotes a *Cartesian coordinate* system. Electric and magnetic fields are vectors whose magnitude and direction can depend on position, as denoted by $\mathbf{E}(\mathbf{r})$ or $\mathbf{B}(\mathbf{r})$. An example of such a field is $\mathbf{E}(\mathbf{r}) = q(\mathbf{r}-\mathbf{r}_0)\big/4\pi\epsilon_0 |\mathbf{r}-\mathbf{r}_0|^3$, which is the static electric field surrounding a point charge located at position $\mathbf{r}_0$. The *absolute-value* brackets indicate the *magnitude* (or length) of the vector given by

$$\begin{aligned} |\mathbf{r}-\mathbf{r}_0| &= \left|(x-x_0)\hat{\mathbf{x}} + (y-y_0)\hat{\mathbf{y}} + (z-z_0)\hat{\mathbf{z}}\right| \\ &= \sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2} \end{aligned} \tag{0.1}$$

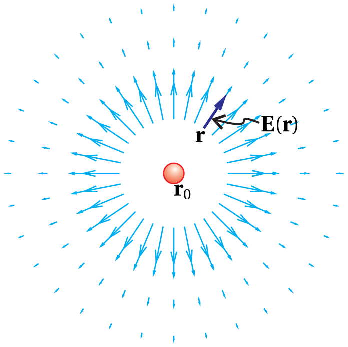


Figure 0.1 The electric field vectors around a point charge.

Example 0.1

Compute the electric field at $\mathbf{r} = \left(2\hat{\mathbf{x}} + 2\hat{\mathbf{y}} + 2\hat{\mathbf{z}}\right)$ Å due to a positive point charge q positioned at $\mathbf{r}_0 = \left(1\hat{\mathbf{x}} + 1\hat{\mathbf{y}} + 2\hat{\mathbf{z}}\right)$ Å.

Solution: As mentioned above, the field is given by $\mathbf{E}(\mathbf{r}) = q\left(\mathbf{r} - \mathbf{r}_0\right)/4\pi\epsilon_0 \left|\mathbf{r} - \mathbf{r}_0\right|^3$. We have

$$\mathbf{r} - \mathbf{r}_0 = \left((2-1)\hat{\mathbf{x}} + (2-1)\hat{\mathbf{y}} + (2-2)\hat{\mathbf{z}}\right) \text{Å} = \left(1\hat{\mathbf{x}} + 1\hat{\mathbf{y}}\right) \text{Å}$$

and

$$|\mathbf{r} - \mathbf{r}_0| = \sqrt{(1)^2 + (1)^2}\,\text{Å} = \sqrt{2}\,\text{Å}$$

The electric field is then

$$\mathbf{E} = \frac{q\left(1\hat{\mathbf{x}} + 1\hat{\mathbf{y}}\right) \text{Å}}{4\pi\epsilon_0\left(\sqrt{2}\,\text{Å}\right)^3}$$

In addition to position, the electric and magnetic fields almost always depend on time in optics problems. For example, a common time-dependent field is $\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 \cos(\mathbf{k}\cdot\mathbf{r} - \omega t)$. The *dot product* $\mathbf{k}\cdot\mathbf{r}$ is an example of *vector multiplication*, and signifies the following operation:

$$\begin{aligned} \mathbf{k}\cdot\mathbf{r} &= \left(k_x\hat{\mathbf{x}} + k_y\hat{\mathbf{y}} + k_z\hat{\mathbf{z}}\right)\cdot\left(x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}\right) \\ &= k_x x + k_y y + k_z z \\ &= |\mathbf{k}||\mathbf{r}|\cos\phi \end{aligned} \tag{0.2}$$

where ϕ is the angle between the vectors $\mathbf{k}$ and $\mathbf{r}$.

Proof of the final line of (0.2)

Consider the plane that contains the two vectors $\mathbf{k}$ and $\mathbf{r}$. Call it the $x'y'$-plane. In this coordinate system, the two vectors can be written as $\mathbf{k} = k\cos\theta\hat{\mathbf{x}}' + k\sin\theta\hat{\mathbf{y}}'$ and $\mathbf{r} = r\cos\alpha\hat{\mathbf{x}}' + r\sin\alpha\hat{\mathbf{y}}'$, where θ and α are the respective angles that the two vectors make with the x'-axis. The dot product gives $\mathbf{k}\cdot\mathbf{r} = kr(\cos\theta\cos\alpha + \sin\theta\sin\alpha)$. This simplifies to $\mathbf{k}\cdot\mathbf{r} = kr\cos\phi$ (see (0.13)), where $\phi \equiv \theta - \alpha$ is the angle between the vectors. Thus, the dot product between two vectors is the product of the magnitudes of the vectors times the cosine of the angle between them.

Another type of vector multiplication is the *cross product*, which is accomplished in the following manner:[1]

$$\begin{aligned} \mathbf{E}\times\mathbf{B} &= \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ E_x & E_y & E_z \\ B_x & B_y & B_z \end{vmatrix} \\ &= \left(E_yB_z - E_zB_y\right)\hat{\mathbf{x}} - \left(E_xB_z - E_zB_x\right)\hat{\mathbf{y}} + \left(E_xB_y - E_yB_x\right)\hat{\mathbf{z}} \end{aligned} \tag{0.3}$$

---

[1] The use of the determinant to generate the cross product is merely a convenient device for remembering its form.

Note that the cross product results in a vector, whereas the dot product mentioned above results in a scalar (i.e. a number with appropriate units). The resultant cross-product vector is always perpendicular to the two vectors that are cross-multiplied. If the fingers on your right hand curl from the first vector towards the second, your thumb will point in the direction of the result. The magnitude of the result equals the product of the magnitudes of the constituent vectors times the sine of the angle between them.

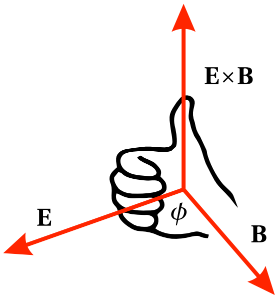


Figure 0.2 Right-hand rule for cross product.

Proof of cross-product properties

We label the plane containing **E** and **B** the $x'y'$-plane. In this coordinate system, the two vectors can be written as $\mathbf{E} = E\cos\theta\hat{\mathbf{x}}' + E\sin\theta\hat{\mathbf{y}}'$ and $\mathbf{B} = B\cos\alpha\hat{\mathbf{x}}' + B\sin\alpha\hat{\mathbf{y}}'$, where θ and α are the respective angles that the two vectors make with the x'-axis. The cross product, according to (0.3), gives $\mathbf{E}\times\mathbf{B} = EB(\cos\theta\sin\alpha - \sin\theta\cos\alpha)\hat{\mathbf{z}}'$. This simplifies to $\mathbf{E}\times\mathbf{B} = EB\sin\phi\hat{\mathbf{z}}'$ (see (0.14)), where $\phi \equiv \alpha - \theta$ is the angle between the vectors. The vectors **E** and **B**, which both lie in the $x'y'$-plane, are both perpendicular to z'. If $0 < \theta - \alpha < \pi$, the result $\mathbf{E}\times\mathbf{B}$ points in the positive z' direction, which is consistent with the right-hand rule.

We will use several multidimensional derivatives in our study of optics, namely the *gradient*, the *divergence*, and the *curl*.[2] In Cartesian coordinates, the gradient of a scalar function is given by

$$\nabla f\left(x,y,z\right) = \frac{\partial f}{\partial x}\hat{\mathbf{x}} + \frac{\partial f}{\partial y}\hat{\mathbf{y}} + \frac{\partial f}{\partial z}\hat{\mathbf{z}} \tag{0.4}$$

the divergence, which applies to vector functions, is given by

$$\nabla\cdot\mathbf{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \tag{0.5}$$

and the curl, which also applies to vector functions, is given by

$$\begin{aligned}\nabla\times\mathbf{E} &= \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ E_x & E_y & E_z \end{vmatrix} \\ &= \left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right)\hat{\mathbf{x}} - \left(\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z}\right)\hat{\mathbf{y}} + \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right)\hat{\mathbf{z}}\end{aligned} \tag{0.6}$$

Example 0.2

Derive the gradient (0.4) in *cylindrical coordinates* defined by the transformations $x = \rho\cos\phi$ and $y = \rho\sin\phi$. (The coordinate z remains unchanged.)

---

[2] See M. R. Spiegel, *Schaum's Outline of Advanced Mathematics for Engineers and Scientists*, pp. 126-127 (New York: McGraw-Hill 1971).

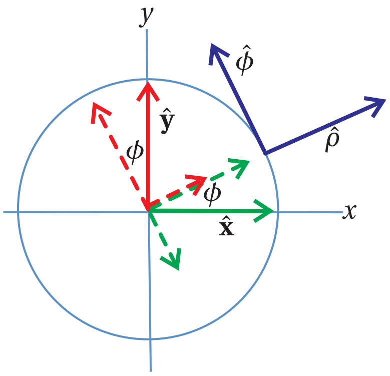


Figure 0.3 The unit vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ may be expressed in terms of components along $\hat{\phi}$ and $\hat{\rho}$ in cylindrical coordinates.

Solution: By inspection of Fig. 0.3, the Cartesian unit vectors may be expressed as

$$\hat{\mathbf{x}} = \cos\phi\hat{\rho} - \sin\phi\hat{\phi} \quad \text{and} \quad \hat{\mathbf{y}} = \sin\phi\hat{\rho} + \cos\phi\hat{\phi}$$

In accordance with the rules of calculus, the needed partial derivatives expressed in terms of the new variables are

$$\frac{\partial}{\partial x} = \left(\frac{\partial\rho}{\partial x}\right)\frac{\partial}{\partial\rho} + \left(\frac{\partial\phi}{\partial x}\right)\frac{\partial}{\partial\phi} \quad \text{and} \quad \frac{\partial}{\partial y} = \left(\frac{\partial\rho}{\partial y}\right)\frac{\partial}{\partial\rho} + \left(\frac{\partial\phi}{\partial y}\right)\frac{\partial}{\partial\phi}$$

Meanwhile, the inverted form of the coordinate transformation is

$$\rho = \sqrt{x^2+y^2} \quad \text{and} \quad \phi = \tan^{-1} y/x$$

from which we obtain the following derivatives:

$$\begin{aligned} \frac{\partial\rho}{\partial x} &= \frac{x}{\sqrt{x^2+y^2}} = \cos\phi & \qquad \frac{\partial\phi}{\partial x} &= -\frac{y}{x^2+y^2} = -\frac{\sin\phi}{\rho} \\ \frac{\partial\rho}{\partial y} &= \frac{y}{\sqrt{x^2+y^2}} = \sin\phi & \qquad \frac{\partial\phi}{\partial y} &= \frac{x}{x^2+y^2} = \frac{\cos\phi}{\rho} \end{aligned}$$

Putting this all together, we arrive at

$$\begin{aligned} \nabla f &= \frac{\partial f}{\partial x}\hat{\mathbf{x}} + \frac{\partial f}{\partial y}\hat{\mathbf{y}} + \frac{\partial f}{\partial z}\hat{\mathbf{z}} \\ &= \left(\cos\phi\frac{\partial f}{\partial\rho} - \frac{\sin\phi}{\rho}\frac{\partial f}{\partial\phi}\right)(\cos\phi\hat{\rho} - \sin\phi\hat{\phi}) \\ &\quad + \left(\sin\phi\frac{\partial f}{\partial\rho} + \frac{\cos\phi}{\rho}\frac{\partial f}{\partial\phi}\right)(\sin\phi\hat{\rho} + \cos\phi\hat{\phi}) + \frac{\partial f}{\partial z}\hat{\mathbf{z}} \\ &= \frac{\partial f}{\partial\rho}\hat{\rho} + \frac{1}{\rho}\frac{\partial f}{\partial\phi}\hat{\phi} + \frac{\partial f}{\partial z}\hat{\mathbf{z}} \end{aligned}$$

where we have used $\cos^2\phi + \sin^2\phi = 1$ (see Ex. 0.4).

We will sometimes need a multidimensional second derivative called the *Laplacian*. When applied to a scalar function, it is defined as the divergence of a gradient:

$$\nabla^2 f(x,y,z) \equiv \nabla\cdot[\nabla f(x,y,z)] \tag{0.7}$$

In Cartesian coordinates, this reduces to

$$\nabla^2 f(x,y,z) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} \tag{0.8}$$

The Laplacian applied to a scalar gives a result that is also a scalar. In Cartesian coordinates, we deal with vector functions by applying the Laplacian to the scalar function attached to each unit vector:

$$\nabla^2\mathbf{E} = \left(\frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2}\right)\hat{\mathbf{x}} + \left(\frac{\partial^2 E_y}{\partial x^2} + \frac{\partial^2 E_y}{\partial y^2} + \frac{\partial^2 E_y}{\partial z^2}\right)\hat{\mathbf{y}} + \left(\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} + \frac{\partial^2 E_z}{\partial z^2}\right)\hat{\mathbf{z}} \tag{0.9}$$

Pierre-Simon Laplace (1749-1827, French) was born in Normandy, France to a farm laborer. Some wealthy neighbors noticed his unusual abilities and took an interest in his education. Laplace is sometimes revered as the "Newton" of France with contributions to mathematics and astronomy. The Laplacian differential operator as well as Laplace transforms are used widely in applied mathematics. (Wikipedia)

This is possible because each unit vector is a constant in Cartesian coordinates.

The various multidimensional derivatives take on more complicated forms in non-Cartesian coordinates such as cylindrical or spherical. You can derive the Laplacian for these other coordinate systems by changing variables and rewriting the unit vectors starting from the above Cartesian expression. (See Problem 0.10.) Regardless of the coordinate system, the Laplacian for a vector function can be obtained from first derivatives though

$$\nabla^2\mathbf{E} \equiv \nabla(\nabla\cdot\mathbf{E}) - \nabla\times(\nabla\times\mathbf{E}) \tag{0.10}$$

Verification of (0.10) **in Cartesian coordinates**

From (0.6), we have

$$\nabla\times\mathbf{E} = \left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right)\hat{\mathbf{x}} - \left(\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z}\right)\hat{\mathbf{y}} + \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right)\hat{\mathbf{z}}$$

and

$$\begin{aligned}
\nabla\times(\nabla\times\mathbf{E}) &= \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ \left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right) & -\left(\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z}\right) & \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right) \end{vmatrix} \\
&= \left[\frac{\partial}{\partial y}\left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right) + \frac{\partial}{\partial z}\left(\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z}\right)\right]\hat{\mathbf{x}} - \left[\frac{\partial}{\partial x}\left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right) - \frac{\partial}{\partial z}\left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right)\right]\hat{\mathbf{y}} \\
&\quad + \left[-\frac{\partial}{\partial x}\left(\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z}\right) - \frac{\partial}{\partial y}\left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right)\right]\hat{\mathbf{z}}
\end{aligned}$$

After adding and subtracting $\frac{\partial^2 E_x}{\partial x^2}\hat{\mathbf{x}} + \frac{\partial^2 E_y}{\partial y^2}\hat{\mathbf{y}} + \frac{\partial^2 E_z}{\partial z^2}\hat{\mathbf{z}}$ and then rearranging, we get

$$\begin{aligned}
\nabla\times(\nabla\times\mathbf{E}) &= \left[\frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_y}{\partial x\partial y} + \frac{\partial^2 E_z}{\partial x\partial z}\right]\hat{\mathbf{x}} + \left[\frac{\partial^2 E_x}{\partial x\partial y} + \frac{\partial^2 E_y}{\partial y^2} + \frac{\partial^2 E_z}{\partial y\partial z}\right]\hat{\mathbf{y}} + \left[\frac{\partial^2 E_x}{\partial x\partial z} + \frac{\partial^2 E_y}{\partial y\partial z} + \frac{\partial^2 E_z}{\partial z^2}\right]\hat{\mathbf{z}} \\
&\quad - \left[\frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2}\right]\hat{\mathbf{x}} - \left[\frac{\partial^2 E_y}{\partial x^2} + \frac{\partial^2 E_y}{\partial y^2} + \frac{\partial^2 E_y}{\partial z^2}\right]\hat{\mathbf{y}} - \left[\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} + \frac{\partial^2 E_z}{\partial z^2}\right]\hat{\mathbf{z}}
\end{aligned}$$

After some factorization, we obtain

$$\begin{aligned}
\nabla\times(\nabla\times\mathbf{E}) &= \left[\hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}\right]\left[\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}\right] - \left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right]\left[E_x\hat{\mathbf{x}} + E_y\hat{\mathbf{y}} + E_z\hat{\mathbf{z}}\right] \\
&= \nabla(\nabla\cdot\mathbf{E}) - \nabla^2\mathbf{E}
\end{aligned}$$

where on the final line we invoked (0.4), (0.5), and (0.8).

We will also encounter several integral theorems[3] involving vector functions. The *divergence theorem* for a vector function **F** is

$$\oint_S \mathbf{F}\cdot\hat{\mathbf{n}}\,da = \int_V \nabla\cdot\mathbf{F}\,dv \tag{0.11}$$

---

[3] For succinct treatments of the divergence theorem and Stokes' theorem, see M. R. Spiegel, *Schaum's Outline of Advanced Mathematics for Engineers and Scientists*, p. 154 (New York: McGraw-Hill 1971).

The integration on the left-hand side is over the closed surface S, which contains the volume V associated with the integration on the right-hand side. The unit vector $\hat{\mathbf{n}}$ points outward, *normal* to the surface. The divergence theorem is especially useful in connection with Gauss' law, where the left-hand side is interpreted as the number of field lines exiting a closed surface.

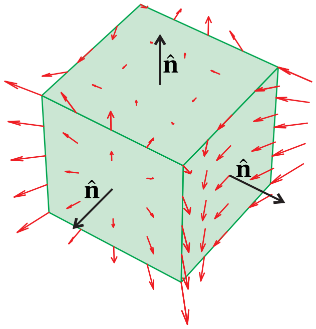


Figure 0.4 The function $\mathbf{F}$ (red arrows) plotted for several points on the surface S.

Example 0.3

Check the divergence theorem (0.11) for the vector function $\mathbf{F}(x,y,z) = y^2\,\hat{\mathbf{x}} + xy\,\hat{\mathbf{y}} + x^2z\,\hat{\mathbf{z}}$. Take as the volume a cube contained by the six planes $x = \pm 1$, $y = \pm 1$, and $z = \pm 1$.

Solution: First, we evaluate the left side of (0.11) for the function:

$$\begin{aligned}\oint_S \mathbf{F}\cdot\hat{\mathbf{n}}\,da &= \int_{-1}^{1}\int_{-1}^{1} dx\,dy\,(x^2z)_{z=1} - \int_{-1}^{1}\int_{-1}^{1} dx\,dy\,(x^2z)_{z=-1} + \int_{-1}^{1}\int_{-1}^{1} dx\,dz\,(xy)_{y=1} \\ &\quad - \int_{-1}^{1}\int_{-1}^{1} dx\,dz\,(xy)_{y=-1} + \int_{-1}^{1}\int_{-1}^{1} dy\,dz\,(y^2)_{x=1} - \int_{-1}^{1}\int_{-1}^{1} dy\,dz\,(y^2)_{x=-1} \\ &= 2\int_{-1}^{1}\int_{-1}^{1} dx\,dy\,x^2 + 2\int_{-1}^{1}\int_{-1}^{1} dx\,dz\,x = 4\,\frac{x^3}{3}\bigg|_{-1}^{1} + 4\,\frac{x^2}{2}\bigg|_{-1}^{1} = \frac{8}{3}\end{aligned}$$

Now we evaluate the right side of (0.11):

$$\int_V \nabla\cdot\mathbf{F}\,dv = \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} dx\,dy\,dz\,[x + x^2] = 4\int_{-1}^{1} dx\,[x + x^2] = 4\left[\frac{x^2}{2} + \frac{x^3}{3}\right]_{-1}^{1} = \frac{8}{3}$$

Another important theorem is *Stokes' theorem:*

$$\int_S (\nabla\times\mathbf{F})\cdot\hat{\mathbf{n}}\,da = \oint_C \mathbf{F}\cdot d\boldsymbol{\ell} \tag{0.12}$$

The integration on the left-hand side is over an open surface S (not enclosing a volume). The integration on the right-hand side is around the edge of the surface. Again, $\hat{\mathbf{n}}$ is a unit vector that always points *normal to the surface*. The vector $d\boldsymbol{\ell}$ points along the curve C that bounds the surface S. If the fingers of your right hand point in the direction of integration around C, then your thumb points in the direction of $\hat{\mathbf{n}}$. Stokes' theorem is especially useful in connection with Ampere's law and Faraday's law. The right-hand side is an integration of a field around a loop.

0.2 Complex Numbers

It is often convenient to represent electromagnetic wave phenomena (i.e. light) as a superposition of sinusoidal functions, each having the form $A\cos(\alpha + \beta)$. The sine function is intrinsically present in this formula via the identity

$$\cos\left(\alpha+\beta\right)=\cos\alpha\cos\beta-\sin\alpha\sin\beta \tag{0.13}$$

This is a good formula to commit to memory, as well as the frequently used identity

$$\sin\left(\alpha+\beta\right)=\sin\alpha\cos\beta+\sin\beta\cos\alpha \tag{0.14}$$

With a basic familiarity with trigonometry, we can approach many optical problems including those involving the addition of multiple waves. However, the manipulation of trigonometric functions via identities such as (0.13) and (0.14) can be cumbersome and tedious. Fortunately, complex-number notation offers an equivalent approach with far less busy work. The modest investment needed to become comfortable with complex notation is definitely worth it; optics problems can become cumbersome enough even with the most efficient methods!

Leonhard Euler (1707-1783, Swiss) was born in Basel, Switzerland. His father, Paul Euler, was friends with the well-known mathematician Johann Bernoulli, who discovered young Euler's great talent for mathematics and tutored him regularly. Euler enrolled at the University of Basel at age thirteen. In 1726 Euler accepted an offer to join the Russian Academy of Sciences in St Petersburg, having unsuccessfully applied for a professorship at the University of Basel. Under the auspices of the Czars (with the exception of 12-year-old Peter II), foreign academicians in the Russian Academy were given considerable freedom to pursue scientific questions with relatively light teaching duties. Euler spent his early career in Russia, his mid career in Berlin, and his later career again in Russia. Euler introduced the concept of a function. He successfully defined logarithms and exponential functions for complex numbers and discovered the connection to trigonometric functions. The special case of Euler's formula $e^{i\pi}+1=0$ has been voted by modern fans of mathematics (including Richard Feynman) as "the Most Beautiful Mathematical Formula Ever" for its single uses of addition, multiplication, exponentiation, equality, and the constants 0, 1, e, i and π. Euler and his wife, Katharina Gsell, were the parents of 13 children, many of whom died in childhood. (Wikipedia)

The convenience of complex-number notation has its origins in *Euler's formula*:

$$e^{i\phi}=\cos\phi+i\sin\phi \tag{0.15}$$

where $i\equiv\sqrt{-1}$ is an *imaginary number*. By inverting Euler's formula (0.15) (and its twin with $\phi\to-\phi$) we can obtain the following representation of the cosine and sine functions:

$$\begin{aligned}\cos\phi&=\frac{e^{i\phi}+e^{-i\phi}}{2},\\ \sin\phi&=\frac{e^{i\phi}-e^{-i\phi}}{2i}\end{aligned} \tag{0.16}$$

Equation (0.16) shows how ordinary sines and cosines are intimately related to *hyperbolic cosines* and *hyperbolic sines*. If ϕ happens to be imaginary such that $\phi=i\gamma$ where γ is real, then we have

$$\begin{aligned}\sin i\gamma&=\frac{e^{-\gamma}-e^{\gamma}}{2i}=i\sinh\gamma\\ \cos i\gamma&=\frac{e^{-\gamma}+e^{\gamma}}{2}=\cosh\gamma\end{aligned} \tag{0.17}$$

Proof of Euler's formula

We can prove Euler's formula using a *Taylor's series* expansion:

$$f(x)=f(x_0)+\frac{1}{1!}(x-x_0)\left.\frac{df}{dx}\right|_{x=x_0}+\frac{1}{2!}(x-x_0)^2\left.\frac{d^2f}{dx^2}\right|_{x=x_0}+\cdots \tag{0.18}$$

By expanding each function appearing in (0.15) in a Taylor's series about the origin we obtain

$$\begin{aligned}\cos\phi&=1-\frac{\phi^2}{2!}+\frac{\phi^4}{4!}-\cdots\\ i\sin\phi&=i\phi-i\frac{\phi^3}{3!}+i\frac{\phi^5}{5!}-\cdots\\ e^{i\phi}&=1+i\phi-\frac{\phi^2}{2!}-i\frac{\phi^3}{3!}+\frac{\phi^4}{4!}+i\frac{\phi^5}{5!}-\cdots\end{aligned} \tag{0.19}$$

Brook Taylor (1685-1731, English) was born in Middlesex, England. He studied at Cambridge as a fellow-commoner earning a bachelor degree in 1709 and a doctoral degree in 1714. Soon thereafter, he developed the branch of mathematics known as calculus of finite differences. He used it to study the movement of vibrating strings. As part of that work, he developed the formula known today as Taylor's theorem, which was under-appreciated until 1772, when French mathematician Lagrange referred to it as "the main foundation of differential calculus." (Wikipedia)

The last line of (0.19) is seen to be the sum of the first two lines, from which Euler's formula directly follows.

Example 0.4

Prove (0.13) and (0.14) as well as $\cos^2\phi + \sin^2\phi = 1$ by taking advantage of (0.16).

Solution: We start with Euler's formula (0.15) for a sum of angles:

$$\begin{aligned}
\cos(\alpha+\beta) + i\sin(\alpha+\beta) &= e^{i(\alpha+\beta)} \\
&= e^{i\alpha}e^{i\beta} \\
&= (\cos\alpha + i\sin\alpha)(\cos\beta + i\sin\beta) \\
&= (\cos\alpha\cos\beta - \sin\alpha\sin\beta) + i(\sin\alpha\cos\beta + \cos\alpha\sin\beta)
\end{aligned}$$

Equating the real parts gives (0.13), and equating the imaginary parts gives (0.14). In the case of $\beta = -\alpha$, we have $1 = \cos^2\alpha + \sin^2\alpha$.

Or,

We start with (0.13). By direct application of (0.16) and some rearranging, we have

$$\begin{aligned}
\cos\alpha\cos\beta - \sin\alpha\sin\beta &= \frac{e^{i\alpha}+e^{-i\alpha}}{2}\frac{e^{i\beta}+e^{-i\beta}}{2} - \frac{e^{i\alpha}-e^{-i\alpha}}{2i}\frac{e^{i\beta}-e^{-i\beta}}{2i} \\
&= \frac{e^{i(\alpha+\beta)}+e^{i(\alpha-\beta)}+e^{-i(\alpha-\beta)}+e^{-i(\alpha+\beta)}}{4} \\
&\quad + \frac{e^{i(\alpha+\beta)}-e^{i(\alpha-\beta)}-e^{-i(\alpha-\beta)}+e^{-i(\alpha+\beta)}}{4} \\
&= \frac{e^{i(\alpha+\beta)}+e^{-i(\alpha+\beta)}}{2} = \cos\left(\alpha+\beta\right)
\end{aligned}$$

We can prove (0.14) using the same technique:

$$\begin{aligned}
\sin\alpha\cos\beta + \sin\beta\cos\alpha &= \frac{e^{i\alpha}-e^{-i\alpha}}{2i}\frac{e^{i\beta}+e^{-i\beta}}{2} + \frac{e^{i\beta}-e^{-i\beta}}{2i}\frac{e^{i\alpha}+e^{-i\alpha}}{2} \\
&= \frac{e^{i(\alpha+\beta)}+e^{i(\alpha-\beta)}-e^{-i(\alpha-\beta)}-e^{-i(\alpha+\beta)}}{4i} \\
&\quad + \frac{e^{i(\alpha+\beta)}-e^{i(\alpha-\beta)}+e^{-i(\alpha-\beta)}-e^{-i(\alpha+\beta)}}{4i} \\
&= \frac{e^{i(\alpha+\beta)}-e^{-i(\alpha+\beta)}}{2i} = \sin\left(\alpha+\beta\right)
\end{aligned}$$

Finally, we compute

$$\begin{aligned}
\cos^2\phi + \sin^2\phi &= \left(\frac{e^{i\phi}+e^{-i\phi}}{2}\right)^2 + \left(\frac{e^{i\phi}-e^{-i\phi}}{2i}\right)^2 \\
&= \frac{e^{2i\phi}+2+e^{-2i\phi}}{4} - \frac{e^{2i\phi}-2+e^{-2i\phi}}{4} = 1
\end{aligned}$$

As was mentioned previously, we will often be interested in waves of the form $A\cos(\alpha+\beta)$. We can use complex notation to represent this wave simply by writing

$$A\cos(\alpha+\beta) = \mathrm{Re}\left\{\tilde{A}e^{i\alpha}\right\} \tag{0.20}$$

where the 'phase' β is conveniently contained within the complex factor $\tilde{A} \equiv Ae^{i\beta}$. The operation $\mathrm{Re}\{\ \}$ means to retain only the *real part* of the argument without regard for the *imaginary part*. As an example, we have $\mathrm{Re}\{1+2i\} = 1$. The formula (0.20) follows directly from Euler's equation (0.15).

It is common (even conventional) to omit the explicit writing of $\mathrm{Re}\{\ \}$. Thus, physicists participate in a conspiracy that $\tilde{A}e^{i\alpha}$ actually means $A\cos(\alpha+\beta)$. This laziness is permissible because it is possible to perform linear operations on $\mathrm{Re}\{f\}$ such as addition, differentiation, or integration while procrastinating the taking of the real part until the end:

$$\begin{aligned}\mathrm{Re}\{f\} + \mathrm{Re}\{g\} &= \mathrm{Re}\{f+g\}\\ \frac{d}{dx}\mathrm{Re}\{f\} &= \mathrm{Re}\left\{\frac{df}{dx}\right\}\\ \int \mathrm{Re}\{f\}\,dx &= \mathrm{Re}\left\{\int f\,dx\right\}\end{aligned} \tag{0.21}$$

As an example, note that $\mathrm{Re}\{1+2i\} + \mathrm{Re}\{3+4i\} = \mathrm{Re}\{(1+2i)+(3+4i)\} = 4$. However, we must be careful when performing other operations such as multiplication. In this case, it is essential to take the real parts before performing the operation. Notice that

$$\mathrm{Re}\{f\} \times \mathrm{Re}\{g\} \neq \mathrm{Re}\{f \times g\} \tag{0.22}$$

As an example, we see $\mathrm{Re}\{1+2i\} \times \mathrm{Re}\{3+4i\} = 3$, but $\mathrm{Re}\{(1+2i)(3+4i)\} = -5$.

When dealing with complex numbers it is often advantageous to transform between a Cartesian representation and a *polar representation*. With the aid of Euler's formula, it is possible to transform any complex number $a+ib$ into the form $\rho e^{i\phi}$, where a, b, ρ, and ϕ are real. From (0.15), the required connection between (ρ,ϕ) and (a,b) is

$$\rho e^{i\phi} = \rho\cos\phi + i\rho\sin\phi = a+ib \tag{0.23}$$

The real and imaginary parts of this equation must separately be equal. Thus, we have

$$\begin{aligned}a &= \rho\cos\phi\\ b &= \rho\sin\phi\end{aligned} \tag{0.24}$$

These equations can be inverted to yield

$$\begin{aligned}\rho &= \sqrt{a^2+b^2}\\ \phi &= \tan^{-1}\frac{b}{a} \qquad (a>0)\end{aligned} \tag{0.25}$$

Gerolamo Cardano (1501-1576, Italian) was the first to introduce the notion of complex numbers (which he called "fictitious") while developing solutions to cubic and quartic equations. He was born in Pavia, Italy, the illegitimate son of a lawyer who was an acquaintance of Leonardo da Vinci. Cardano was fortunate to survive infancy as his father claimed that his mother attempted to abort him and his older siblings all died of the plague. Cardano studied at the University of Pavia and later at Padua. He was known for being eccentric and confrontational, which did not earn him many friends. He supported himself in part as a somewhat successful gambler, but he was often short of money. Cardano also introduced binomial coefficients and the binomial theorem. (Wikipedia)

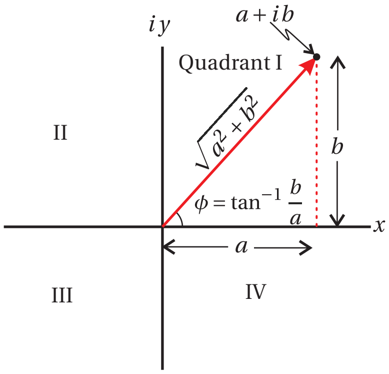


Figure 0.5 A number in the complex plane can be represented either by Cartesian or polar representation.

When $a < 0$, we must adjust ϕ by π since the arctangent has a range only from $-\pi/2$ to $\pi/2$.

The transformations in (0.24) and (0.25) have a clear geometrical interpretation in the *complex plane*, and this makes it easier to remember them. They are just the usual connections between Cartesian and polar coordinates. As seen in Fig. 0.5, ρ is the hypotenuse of a right triangle having legs with lengths a and b, and ϕ is the angle that the hypotenuse makes with the x-axis. Again, you should be careful when a is negative since the arctangent is defined in quadrants I and IV. An easy way to deal with the situation of a negative a is to factor the minus sign out before proceeding (i.e. $a + ib = -(-a - ib)$). Then the transformation is made on $-a - ib$ where $-a$ is positive. The overall minus sign out in front is just carried along unaffected and can be factored back in at the end. Notice that $-\rho e^{i\phi}$ is the same as $\rho e^{i(\phi\pm\pi)}$.

Example 0.5

Write $-3 + 4i$ in polar format.

Solution: We must be careful with the negative real part since it indicates a quadrant (in this case II) outside of the domain of the inverse tangent (quadrants I and IV). Best to factor the negative out and deal with it separately.

$$-3+4i = -(3-4i) = -\sqrt{3^2+(-4)^2}\,e^{i\tan^{-1}\frac{(-4)}{3}} = e^{i\pi}5e^{-i\tan^{-1}\frac{4}{3}} = 5e^{i\left(\pi-\tan^{-1}\frac{4}{3}\right)}$$

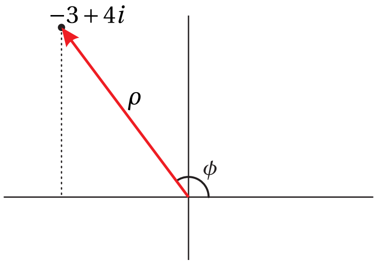


Figure 0.6 Geometric representation of $-3+4i$

Finally, we consider the concept of a *complex conjugate*. The conjugate of a complex number $z = a + ib$ is denoted with an asterisk and amounts to changing the sign on the imaginary part of the number:

$$z^* = (a+ib)^* \equiv a - ib \tag{0.26}$$

The complex conjugate is useful when computing the *absolute value* of a complex number:

$$|z| = \sqrt{z^* z} = \sqrt{(a-ib)(a+ib)} = \sqrt{a^2+b^2} = \rho \tag{0.27}$$

Note that the absolute value of a complex number is the same as its magnitude ρ as defined in (0.25). The complex conjugate is also useful for eliminating complex numbers from the denominator of expressions:

$$\frac{a+ib}{c+id} = \frac{(a+ib)}{(c+id)}\frac{(c-id)}{(c-id)} = \frac{ac+bd+i(bc-ad)}{c^2+d^2} \tag{0.28}$$

No matter how complicated an expression, the complex conjugate is calculated by inserting a minus sign in front of all occurrences of i in the expression, and placing an asterisk on all complex variables in the expression. For example, the complex conjugate of $\rho e^{i\phi}$ is $\rho e^{-i\phi}$ assuming ρ and ϕ are real, as can be seen from Euler's formula (0.15). As another example consider

$$\left[E_0 \exp\{i(kz-\omega t)\}\right]^* = E_0^* \exp\left\{-i\left(k^* z - \omega t\right)\right\} \tag{0.29}$$

assuming z, ω, and t are real, but E_0 and k are complex.

A common way of obtaining the real part of an expression is by adding the complex conjugate and dividing the result by 2:

$$\mathrm{Re}\{z\} = \frac{1}{2}\left(z + z^*\right) \tag{0.30}$$

Notice that the expression for $\cos\phi$ in (0.16) is an example of this formula. Sometimes when a lengthy expression is added to its own complex conjugate, we let "C.C." represent the complex conjugate in order to avoid writing the expression twice.

In optics we sometimes encounter a *complex angle*, such as kz in (0.29). The imaginary part of k governs exponential decay (or growth) when a light wave propagates in an absorptive (or amplifying) medium. Similarly, when we compute the transmission angle for light incident upon a surface beyond the critical angle for total internal reflection, we encounter the arcsine of a number greater than one in an effort to satisfy Snell's law. Even though such an angle does not exist in the physical sense, a complex value for the angle can be found, which satisfies (0.16) and describes evanescent waves.

0.3 Linear Algebra

Throughout this book we will often encounter sets of linear equations. (They are called linear equations because they represent lines in a plane or in space.) Most often, there are just two equations with two variables to solve. The simplest example of such a set of equations is

$$Ax + By = F \qquad \text{and} \qquad Cx + Dy = G \tag{0.31}$$

where x and y are variables. A set of linear equations such as (0.31) can be expressed using matrix notation as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} Ax + By \\ Cx + Dy \end{bmatrix} = \begin{bmatrix} F \\ G \end{bmatrix} \tag{0.32}$$

As seen above, the 2×2 matrix multiplied onto the two-dimensional column vector results in a two-dimensional vector. The elements of rows are multiplied onto elements of the column and summed to create each new element in the result. A matrix can also be multiplied onto another matrix (rows multiplying columns, resulting in a matrix). The order of multiplication is important; matrix multiplication is not commutative.

To solve a matrix equation such as (0.32), we multiply both sides by an *inverse matrix*, which gives

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} \begin{bmatrix} F \\ G \end{bmatrix} \tag{0.33}$$

The inverse matrix has the property that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{0.34}$$

where the right-hand side is called the *identity matrix*. You can easily check that the identity matrix leaves unchanged anything that it multiplies, and so (0.33) simplifies to

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} \begin{bmatrix} F \\ G \end{bmatrix}$$

Once the inverse matrix is found, the matrix multiplication on the right can be performed and the answers for x and y obtained as the upper and lower elements of the result.

The inverse of a 2 × 2 matrix is given by

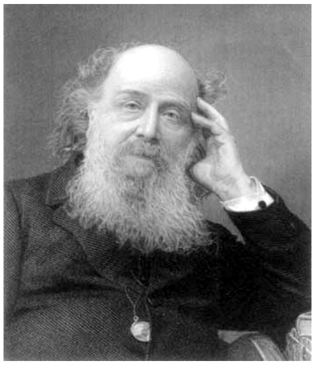
$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \frac{1}{\begin{vmatrix} A & B \\ C & D \end{vmatrix}} \begin{bmatrix} D & -B \\ -C & A \end{bmatrix} \tag{0.35}$$

where

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} \equiv AD - CB$$

is called the *determinant*. We can check that (0.35) is correct by direct substitution:

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} \begin{bmatrix} A & B \\ C & D \end{bmatrix} &= \frac{1}{AD-BC} \begin{bmatrix} D & -B \\ -C & A \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \\ &= \frac{1}{AD-BC} \begin{bmatrix} AD-BC & 0 \\ 0 & AD-BC \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned} \tag{0.36}$$



James Joseph Sylvester (1814-1897, English) made fundamental contributions to matrix theory, invariant theory, number theory, partition theory and combinatorics. He played a leadership role in American mathematics in the later half of the 19th century as a professor at the Johns Hopkins University and as founder of the American Journal of Mathematics. (Wikipedia)

The above review of linear algebra is very basic. In contrast, we next discuss *Sylvester's theorem*, which you probably have not previously encountered. Sylvester's theorem is useful when multiplying the same 2 × 2 matrix (with a determinate of unity) together many times (i.e. raising the matrix to a power). This situation occurs when modeling periodic multilayer mirror coatings or when considering light rays trapped in a laser cavity as they reflect many times.

Sylvester's Theorem:[4] If the determinant of a 2×2 matrix is one, (i.e. $AD - BC = 1$) then

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{N} = \frac{1}{\sin\theta} \begin{bmatrix} A\sin N\theta - \sin(N-1)\theta & B\sin N\theta \\ C\sin N\theta & D\sin N\theta - \sin(N-1)\theta \end{bmatrix} \tag{0.37}$$

---

[4] The theorem presented here is a specific case. See A. A. Tovar and L. W. Casperson, "Generalized Sylvester theorems for periodic applications in matrix optics," J. Opt. Soc. Am. A **12**, 578-590 (1995).

where

$$\cos\theta = \frac{1}{2}(A+D) \tag{0.38}$$

Proof of Sylvester's theorem by induction

When $N = 1$, the equation is seen to be correct by direct substitution. Next we assume that the theorem holds for arbitrary N, and we check to see if it holds for $N+1$:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{N+1} = \frac{1}{\sin\theta}\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} A\sin N\theta - \sin(N-1)\theta & B\sin N\theta \\ C\sin N\theta & D\sin N\theta - \sin(N-1)\theta \end{bmatrix}$$

$$= \frac{1}{\sin\theta}\begin{bmatrix} \left(A^2+BC\right)\sin N\theta - A\sin(N-1)\theta & (AB+BD)\sin N\theta - B\sin(N-1)\theta \\ (AC+CD)\sin N\theta - C\sin(N-1)\theta & \left(D^2+BC\right)\sin N\theta - D\sin(N-1)\theta \end{bmatrix}$$

Now we inject the condition $AD - BC = 1$ into the diagonal elements and obtain

$$\frac{1}{\sin\theta}\begin{bmatrix} \left(A^2+AD-1\right)\sin N\theta - A\sin(N-1)\theta & B\left[(A+D)\sin N\theta - \sin(N-1)\theta\right] \\ C\left[(A+D)\sin N\theta - \sin(N-1)\theta\right] & \left(D^2+AD-1\right)\sin N\theta - D\sin(N-1)\theta \end{bmatrix}$$

and then

$$\frac{1}{\sin\theta}\begin{bmatrix} A\left[(A+D)\sin N\theta - \sin(N-1)\theta\right] - \sin N\theta & B\left[(A+D)\sin N\theta - \sin(N-1)\theta\right] \\ C\left[(A+D)\sin N\theta - \sin(N-1)\theta\right] & D\left[(A+D)\sin N\theta - \sin(N-1)\theta\right] - \sin N\theta \end{bmatrix}$$

In each matrix element, the expression

$$(A+D)\sin N\theta = 2\cos\theta\sin N\theta = \sin(N+1)\theta + \sin(N-1)\theta \tag{0.39}$$

occurs, which we have rearranged using $\cos\theta = \frac{1}{2}(A+D)$ while twice invoking (0.14). The result is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{N+1} = \frac{1}{\sin\theta}\begin{bmatrix} A\sin(N+1)\theta - \sin N\theta & B\sin(N+1)\theta \\ C\sin(N+1)\theta & D\sin(N+1)\theta - \sin N\theta \end{bmatrix}$$

which completes the proof.

0.4 Fourier Theory

In the study of optics, it is common to decompose complicated light fields into superpositions of pure sinusoidal waves. This is called Fourier analysis.[5] This is important since individual sine waves tend to move differently through optical systems (say, a piece of glass with frequency-dependent index). After propagation through a system, we can also reassemble sinusoidal waves to see the effect on the overall waveform. In fact, it will be possible to work simultaneously with infinitely many sinusoidal waves, where the frequencies comprising a light field are spread continuously over a range. Fourier transforms are also helpful for diffraction problems where many waves (all with the same frequency) interfere spatially.

---

[5] See Murray R. Spiegel, *Schaum's Outline of Advanced Mathematics for Engineers and Scientists,* Chaps. 7-8 (New York: McGraw-Hill 1971).

We begin with a derivation of the *Fourier integral theorem.* As asserted by Fourier, a *periodic* function can be represented in terms of sines and cosines in the following manner:

$$f(t) = \sum_{n=0}^{\infty} a_n \cos(n\Delta\omega t) + b_n \sin(n\Delta\omega t) \tag{0.40}$$

This is called a *Fourier expansion.* It is similar in idea to a Taylor's series (0.18), which rewrites a function as a polynomial. In both cases, the goal is to represent one function in terms of a linear combination of other functions (requiring a complete basis set). In a Taylor's series the basis functions are polynomials and in a Fourier expansion the basis functions are sines and cosines with various frequencies (multiples of a fundamental frequency).

Joseph Fourier (1768-1830, French) was born to a tailor in Auxerre, France. He was orphaned at age eight. Because of his humble background, which closed some doors to his education and career, he became a prominent supporter of the French Revolution. He was rewarded by an appointment to a position in the École Polytechnique. In 1798, participated in Napoleon's expedition to Egypt and served as governor over lower Egypt for a time. Fourier made significant contributions to the study of heat transfer and vibrations (presented in 1822), and it was in this context that he asserted that functions could be represented as a series of sine waves. (Wikipedia)

By inspection, we see that all terms in (0.40) repeat with a maximum period of $2\pi/\Delta\omega$. In other words, a Fourier series is good for functions where $f(t) = f(t + 2\pi/\Delta\omega)$. The expansion (0.40) is useful even if $f(t)$ is complex, requiring a_n and b_n to be complex.

Using (0.16), we can rewrite the sines and cosines in the expansion (0.40) as

$$\begin{aligned} f(t) &= \sum_{n=0}^{\infty} a_n \frac{e^{in\Delta\omega t} + e^{-in\Delta\omega t}}{2} + b_n \frac{e^{in\Delta\omega t} - e^{-in\Delta\omega t}}{2i} \\ &= a_0 + \sum_{n=1}^{\infty} \frac{a_n - ib_n}{2} e^{in\Delta\omega t} + \sum_{n=1}^{\infty} \frac{a_n + ib_n}{2} e^{-in\Delta\omega t} \end{aligned} \tag{0.41}$$

or more simply as

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{-in\Delta\omega t} \tag{0.42}$$

where

$$\begin{aligned} c_{n<0} &\equiv \frac{a_{-n} - ib_{-n}}{2} \\ c_{n>0} &\equiv \frac{a_n + ib_n}{2} \\ c_0 &\equiv a_0 \end{aligned} \tag{0.43}$$

Notice that if $c_{-n} = c_n^*$ for all n, then $f(t)$ is real (i.e. real a_n and b_n); otherwise $f(t)$ is complex. The real parts of the c_n coefficients are connected with the cosine terms in (0.40), and the imaginary parts of the c_n coefficients are connected with the sine terms in (0.40).

Given a known function $f(t)$, we can compute the various coefficients c_n. There is a trick for figuring out how to do this. We multiply both sides of (0.42) by

$e^{im\Delta\omega t}$, where m is an integer, and integrate over the function period $2\pi/\Delta\omega$:

$$\begin{aligned}\int_{-\pi/\Delta\omega}^{\pi/\Delta\omega} f(t)e^{im\Delta\omega t}dt &= \sum_{n=-\infty}^{\infty} c_n \int_{-\pi/\Delta\omega}^{\pi/\Delta\omega} e^{i(m-n)\Delta\omega t}dt \\ &= \sum_{n=-\infty}^{\infty} c_n \left[\frac{e^{i(m-n)\Delta\omega t}}{i(m-n)\Delta\omega}\right]_{-\pi/\Delta\omega}^{\pi/\Delta\omega} \\ &= \sum_{n=-\infty}^{\infty} \frac{2\pi c_n}{\Delta\omega}\left[\frac{e^{i(m-n)\pi}-e^{-i(m-n)\pi}}{2i(m-n)\pi}\right] \\ &= \sum_{n=-\infty}^{\infty} \frac{2\pi c_n}{\Delta\omega}\frac{\sin[(m-n)\pi]}{(m-n)\pi}\end{aligned} \tag{0.44}$$

The function $\sin[(m-n)\pi]/[(m-n)\pi]$ is equal to zero for all $n \neq m$, and it is equal to one when $n = m$ (to see this, use L'Hospital's rule on the zero-over-zero situation, or just go back and reperform the above integral for $n = m$). Thus, only one term contributes to the summation in (0.44). We now have

$$c_m = \frac{\Delta\omega}{2\pi}\int_{-\pi/\Delta\omega}^{\pi/\Delta\omega} f(t)e^{im\Delta\omega t}dt \tag{0.45}$$

from which the coefficients c_n can be computed, given a function $f(t)$. (Note that m is a dummy index so we can change it back to n if we like.)

This completes the circle. If we know the function $f(t)$, we can find the coefficients c_n via (0.45), and, if we know the coefficients c_n, we can generate the function $f(t)$ via (0.42). If we are feeling a bit silly, we might combine these into a single identity:

$$f(t) = \sum_{n=-\infty}^{\infty}\left[\frac{\Delta\omega}{2\pi}\int_{-\pi/\Delta\omega}^{\pi/\Delta\omega} f(t')e^{in\Delta\omega t'}dt'\right]e^{-in\Delta\omega t} \tag{0.46}$$

We start with a function $f(t)$ followed by a lot of computation and obtain the function back again! (This is not quite as foolish as it first appears, as we will discuss later.)

As mentioned above, Fourier expansions represent functions $f(t)$ that are periodic over the interval $2\pi/\Delta\omega$. This is disappointing since many optical waveforms do not repeat (e.g. a single short laser pulse). Nevertheless, we can represent a function $f(t)$ that is not periodic if we let the period $2\pi/\Delta\omega$ become infinitely long. In other words, we can accommodate nonperiodic functions if we take the limit as $\Delta\omega$ goes to zero so that the spacing of terms in the series becomes very fine. Applying this limit to (0.46) we obtain

$$f(t) = \frac{1}{2\pi}\lim_{\Delta\omega\to 0}\sum_{n=-\infty}^{\infty}\left[e^{-in\Delta\omega t}\int_{-\infty}^{\infty} f\left(t'\right)e^{in\Delta\omega t'}dt'\right]\Delta\omega \tag{0.47}$$

At this point, a brief review of the definition of an integral is helpful to better understand the next step that we shall administer to (0.47).

Changing the summation in (0.47) **over to an integral**

Recall that an integral is really a summation of rectangles under a curve with finely spaced steps:

$$\begin{aligned}\int_a^b g(\omega)\,d\omega &\equiv \lim_{\Delta\omega\to 0}\sum_{n=0}^{\frac{b-a}{\Delta\omega}} g(a+n\Delta\omega)\,\Delta\omega \\ &= \lim_{\Delta\omega\to 0}\sum_{n=-\frac{b-a}{2\Delta\omega}}^{\frac{b-a}{2\Delta\omega}} g\left(\frac{a+b}{2}+n\Delta\omega\right)\Delta\omega\end{aligned}\tag{0.48}$$

The final expression has been manipulated so that the index ranges through both negative and positive numbers. If we set $a=-b$ and take the limit $b\to\infty$, then the above expression becomes

$$\int_{-\infty}^{\infty} g(\omega)\,d\omega = \lim_{\Delta\omega\to 0}\sum_{n=-\infty}^{\infty} g(n\Delta\omega)\,\Delta\omega \tag{0.49}$$

This concludes our short review of calculus.

Now, (0.47) has the same form as (0.49) if $g(n\Delta\omega)$ represents everything in the square brackets of (0.47). The result is the *Fourier integral theorem*:

$$f(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega t}\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f\left(t'\right)e^{i\omega t'}\,dt'\right]d\omega \tag{0.50}$$

The piece in brackets is called the *Fourier transform*, and the rest of the operation is called the *inverse Fourier transform*. The Fourier integral theorem (0.50) is often written with the following (potentially confusing) notation:

$$\begin{aligned}f(\omega) &\equiv \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)e^{i\omega t}\,dt \\ f(t) &\equiv \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(\omega)\,e^{-i\omega t}\,d\omega\end{aligned}\tag{0.51}$$

The transform and inverse transform are also sometimes written as $f(\omega)\equiv \mathscr{F}\left\{f(t)\right\}$ and $f(t)\equiv\mathscr{F}^{-1}\left\{f(\omega)\right\}$. Note that the functions $f(t)$ and $f(\omega)$ are entirely different, even taking on different units (i.e. the latter having extra units of per frequency). The two functions are distinguished by their arguments, which also have different units (e.g. time vs. frequency). Nevertheless, it is customary to use the same letter to denote either function since they form a *transform pair*.

You should be aware that it is arbitrary which of the expressions in (0.51) is called the transform and which is called the inverse transform. In other words, the signs in the exponents of (0.51) may be interchanged (and this convention varies in published works!). Also, the factor 2π may be placed on either the transform or the inverse transform, or divided equally between the two as has been done here.

> **Example 0.6**
>
> Compute the Fourier transform of $E(t) = E_0 e^{-t^2/2T^2} e^{-i\omega_0 t}$ followed by the inverse Fourier transform.
>
> **Solution:** According to (0.51), the Fourier transform is
>
> $$E(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(E_0 e^{-t^2/2T^2} e^{-i\omega_0 t} \right) e^{i\omega t} dt = \frac{E_0}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2T^2 + i(\omega-\omega_0)t} dt$$
>
> The integration can be performed with the help of (0.55), which yields
>
> $$E(\omega) = \frac{E_0}{\sqrt{2\pi}} \sqrt{\frac{\pi}{1/2T^2}} e^{-\frac{(\omega-\omega_0)^2}{4\left(1/2T^2\right)}} = T E_0 e^{-T^2(\omega-\omega_0)^2/2}$$
>
> Similarly, the inverse Fourier transform of the above function is
>
> $$E(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(T E_0 e^{-T^2(\omega-\omega_0)^2/2} \right) e^{-i\omega t} d\omega = \frac{T E_0}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{T^2}{2}\omega^2 + (T^2\omega_0 - it)\omega - \frac{T^2}{2}\omega_0^2} d\omega$$
>
> where again we use (0.55) to obtain
>
> $$E(t) = \frac{T E_0}{\sqrt{2\pi}} \sqrt{\frac{\pi}{T^2/2}} e^{\frac{\left(T^2\omega_0 - it\right)^2}{4\left(T^2/2\right)} - \frac{T^2}{2}\omega_0^2} = E_0 e^{-t^2/2T^2 - i\omega_0 t}$$
>
> which brings us back to where we started.

As was previously mentioned, it would seem rather pointless to perform a Fourier transform on the function $f(t)$ followed by an inverse Fourier transform, just to end up with $f(t)$ again. Instead, we will typically apply a frequency-dependent effect on $f(\omega)$ before performing the inverse Fourier transform. In this case, the final function will be different from $f(t)$. Keep in mind that $f(\omega)$ is the continuous analog of the discrete coefficients c_n (or the a_n and b_n). The real part of $f(\omega)$ indicates the amplitudes of the cosine waves necessary to construct the function $f(t)$. The imaginary part of $f(\omega)$ indicates the amplitudes of the sine waves necessary to construct the function $f(t)$.

Finally, we comment on the *Dirac delta function*,[6] which is defined indirectly through

$$f(t) = \int_{-\infty}^{\infty} f\left(t'\right) \delta\left(t' - t\right) dt' \tag{0.52}$$

---

[6] See G. B. Arfken and H. J. Weber, *Mathematical Methods for Physicists* 6th ed., Sect. 1.15 (San Diego: Elsevier Academic Press 2005).

The delta function $\delta\left(t'-t\right)$ is zero everywhere except at $t'=t$ where it is infinite in such a way as to make the integral take on the value of the function $f(t)$. (You can think of $\delta\left(t'-t\right)dt'$ as an infinitely tall and infinitely thin rectangle centered at $t'=t$ with an area unity.) The integral only pays attention to the value of $f\left(t'\right)$ at the point $t'=t$.

A remarkable attribute of the delta function can be seen from the Fourier integral theorem. After rearranging the order of integration, the Fourier integral theorem (0.50) can be written as

$$f(t)=\int_{-\infty}^{\infty} f\left(t'\right)\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega\left(t'-t\right)}d\omega\right]dt' \tag{0.53}$$

A comparison between (0.52) and (0.53) shows that you may write the delta function as a uniform superposition of all frequency components:

$$\delta\left(t'-t\right)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega\left(t'-t\right)}\,d\omega \tag{0.54}$$

Example 0.7

Use (0.54) to prove *Parseval's theorem*:[7]

$$\int_{-\infty}^{\infty}\left|f(\omega)\right|^2 d\omega=\int_{-\infty}^{\infty}\left|f(t)\right|^2 dt$$

which comes up often in the study of optics.

Solution:

$$\begin{aligned}\int_{-\infty}^{\infty}\left|f(\omega)\right|^2 d\omega &= \int_{-\infty}^{\infty} f(\omega)f^*(\omega)\,d\omega\\ &= \int_{-\infty}^{\infty}\left\{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)\,e^{i\omega t}dt\right\}\left\{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f^*\left(t'\right)e^{-i\omega t'}dt'\right\}d\omega\end{aligned}$$

The order of integration can be changed, and with the aid of (0.54) we get

$$\begin{aligned}\int_{-\infty}^{\infty}\left|f(\omega)\right|^2 d\omega &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t)f^*\left(-t'\right)\left\{\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega\left(t'-(-t)\right)}d\omega\right\}dt\,dt'\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t)f^*\left(-t'\right)\delta\left(t'-(-t)\right)dt\,dt'\\ &= \int_{-\infty}^{\infty} f(t)f^*(t)\,dt=\int_{-\infty}^{\infty}\left|f(t)\right|^2 dt\end{aligned}$$

---

[7] For a more general version of the relation, see G. B. Arfken and H. J. Weber, *Mathematical Methods for Physicists* 6th ed., Sect. 15.5 (San Diego: Elsevier Academic Press 2005).

Appendix 0.A Table of Integrals and Sums

The following formulas are useful for various problems encountered in the text.

$$\int_{-\infty}^{\infty} e^{-ax^2+bx+c}\,dx = \sqrt{\frac{\pi}{a}}\,e^{\frac{b^2}{4a}+c} \qquad (\mathrm{Re}\{a\}>0) \tag{0.55}$$

$$\int_{-\infty}^{\infty} \frac{e^{iax}}{1+x^2/b^2}\,dx = \pi\,|b|\,e^{-|ab|} \tag{0.56}$$

$$\int_{0}^{2\pi} e^{\pm ia\cos(\theta-\theta')}\,d\theta = 2\pi J_0(a) \tag{0.57}$$

$$\int_{0}^{a} J_0(bx)\,x\,dx = \frac{a}{b}J_1(ab) \tag{0.58}$$

$$\int_{0}^{\infty} e^{-ax^2} J_0(bx)\,x\,dx = \frac{e^{-b^2/4a}}{2a} \tag{0.59}$$

$$\int_{0}^{\infty} \frac{\sin^2(ax)}{(ax)^2}\,dx = \frac{\pi}{2a} \tag{0.60}$$

$$\int \frac{dy}{\left[y^2+c\right]^{3/2}} = \frac{y}{c\sqrt{y^2+c}} \tag{0.61}$$

$$\int \frac{dx}{x\sqrt{x^2-c}} = -\frac{1}{\sqrt{c}}\sin^{-1}\frac{\sqrt{c}}{|x|} \tag{0.62}$$

$$\int_{0}^{\pi} \sin(ax)\sin(bx)\,dx = \int_{0}^{\pi} \cos(ax)\cos(bx)\,dx = \frac{\pi}{2}\delta_{ab} \qquad (a, b \text{ integer}) \tag{0.63}$$

$$\sum_{n=0}^{N} r^n = \frac{1-r^{N+1}}{1-r} \tag{0.64}$$

$$\sum_{n=1}^{N} r^n = \frac{r(1-r^N)}{1-r} \tag{0.65}$$

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r} \qquad (r<1) \tag{0.66}$$

Exercises

Exercises for 0.1 Vector Calculus

P0.1 Let $\mathbf{r} = \left(\hat{\mathbf{x}} + 2\hat{\mathbf{y}} - 3\hat{\mathbf{z}}\right)$ m and $\mathbf{r}_0 = \left(-\hat{\mathbf{x}} + 3\hat{\mathbf{y}} + 2\hat{\mathbf{z}}\right)$ m.

(a) Find the magnitude of $\mathbf{r}$, or in other words r.

(b) Find $\mathbf{r} - \mathbf{r}_0$.

(c) Find the angle between $\mathbf{r}$ and $\mathbf{r}_0$.

Answer: (a) $r = \sqrt{14}$ m; (c) 94°.

P0.2 Use the dot product (0.2) to show that the cross product $\mathbf{E} \times \mathbf{B}$ is perpendicular to $\mathbf{E}$ and to $\mathbf{B}$.

P0.3 Verify the "BAC-CAB" rule: $\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B}\,(\mathbf{A} \cdot \mathbf{C}) - \mathbf{C}\,(\mathbf{A} \cdot \mathbf{B})$.

P0.4 Prove the following identity:

$$\nabla_{\mathbf{r}} \frac{1}{|\mathbf{r} - \mathbf{r}'|} = -\frac{\left(\mathbf{r} - \mathbf{r}'\right)}{|\mathbf{r} - \mathbf{r}'|^3},$$

where $\nabla_{\mathbf{r}}$ operates only on $\mathbf{r}$, treating $\mathbf{r}'$ as a constant vector.

P0.5 Prove that $\nabla_{\mathbf{r}} \cdot \frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}$ is zero, except at $\mathbf{r} = \mathbf{r}'$ where a singularity situation occurs. As in P0.4, $\nabla_{\mathbf{r}}$ operates only on $\mathbf{r}$, treating $\mathbf{r}'$ as a constant vector.

P0.6 Verify $\nabla \cdot (\nabla \times \mathbf{f}) = 0$ for any vector function $\mathbf{f}$.

P0.7 Verify $\nabla \times \left(\mathbf{f} \times \mathbf{g}\right) = \mathbf{f}\left(\nabla \cdot \mathbf{g}\right) - \mathbf{g}\,(\nabla \cdot \mathbf{f}) + \left(\mathbf{g} \cdot \nabla\right)\mathbf{f} - (\mathbf{f} \cdot \nabla)\,\mathbf{g}$.

P0.8 Verify $\nabla \cdot \left(\mathbf{f} \times \mathbf{g}\right) = \mathbf{g} \cdot (\nabla \times \mathbf{f}) - \mathbf{f} \cdot \left(\nabla \times \mathbf{g}\right)$.

P0.9 Verify the following identities:

(a) $\nabla \cdot \left(g\mathbf{f}\right) = \mathbf{f} \cdot \nabla g + g \nabla \cdot \mathbf{f}$

(b) $\nabla \times \left(g\mathbf{f}\right) = \left(\nabla g\right) \times \mathbf{f} + g \nabla \times \mathbf{f}$.

(c) $\nabla \times \left(g\mathbf{f}\right) = g\,(\nabla \times \mathbf{f}) - \mathbf{f} \times \nabla g$

P0.10 Show that the Laplacian in cylindrical coordinates can be written as

$$\nabla^2 = \frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho \frac{\partial}{\partial \rho}\right) + \frac{1}{\rho^2}\frac{\partial^2}{\partial \phi^2} + \frac{\partial^2}{\partial z^2}$$

Solution: (Partial)

Continuing with the approach in Example 0.2, we have

$$\begin{aligned}\frac{\partial^2 f}{\partial x^2} &= \left(\frac{\partial^2 \rho}{\partial x^2}\right)\frac{\partial f}{\partial \rho} + \frac{\partial \rho}{\partial x}\frac{\partial}{\partial \rho}\frac{\partial f}{\partial x} + \left(\frac{\partial^2 \phi}{\partial x^2}\right)\frac{\partial f}{\partial \phi} + \frac{\partial \phi}{\partial x}\frac{\partial}{\partial \phi}\frac{\partial f}{\partial x} \\ &= \left(\frac{\partial^2 \rho}{\partial x^2}\right)\frac{\partial f}{\partial \rho} + \frac{\partial \rho}{\partial x}\frac{\partial}{\partial \rho}\left[\left(\frac{\partial \rho}{\partial x}\right)\frac{\partial f}{\partial \rho} + \left(\frac{\partial \phi}{\partial x}\right)\frac{\partial f}{\partial \phi}\right] + \left(\frac{\partial^2 \phi}{\partial x^2}\right)\frac{\partial f}{\partial \phi} + \frac{\partial \phi}{\partial x}\frac{\partial}{\partial \phi}\left[\left(\frac{\partial \rho}{\partial x}\right)\frac{\partial f}{\partial \rho} + \left(\frac{\partial \phi}{\partial x}\right)\frac{\partial f}{\partial \phi}\right]\end{aligned}$$

and

$$
\begin{aligned}
\nabla^2 f &= \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} \\
&= \left(\frac{\partial^2 \rho}{\partial x^2} + \frac{\partial^2 \rho}{\partial y^2}\right)\frac{\partial f}{\partial \rho} + \left(\left(\frac{\partial \rho}{\partial x}\right)^2 + \left(\frac{\partial \rho}{\partial y}\right)^2\right)\frac{\partial^2 f}{\partial \rho^2} + 2\left[\left(\frac{\partial \phi}{\partial x}\right)\left(\frac{\partial \rho}{\partial x}\right) + \left(\frac{\partial \phi}{\partial y}\right)\left(\frac{\partial \rho}{\partial y}\right)\right]\frac{\partial^2 f}{\partial \phi \partial \rho} \\
&\quad + \left[\left(\frac{\partial^2 \phi}{\partial x^2}\right) + \left(\frac{\partial^2 \phi}{\partial y^2}\right)\right]\frac{\partial f}{\partial \phi} + \left[\left(\frac{\partial \phi}{\partial x}\right)^2 + \left(\frac{\partial \phi}{\partial y}\right)^2\right]\frac{\partial^2 f}{\partial \phi^2} + \frac{\partial^2 f}{\partial z^2}
\end{aligned}
$$

The needed first derivatives are given in Example 0.2. The needed second derivatives are

$$
\begin{aligned}
\frac{\partial^2 \rho}{\partial x^2} &= \frac{1}{\sqrt{x^2+y^2}} - \frac{x^2}{\left(x^2+y^2\right)^{3/2}} = \frac{\sin^2 \phi}{\rho} \\
\frac{\partial^2 \phi}{\partial x^2} &= \frac{2xy}{\left(x^2+y^2\right)^2} = \frac{2\sin\phi\cos\phi}{\rho^2} \\
\frac{\partial^2 \rho}{\partial y^2} &= \frac{1}{\sqrt{x^2+y^2}} - \frac{y^2}{\left(x^2+y^2\right)^{3/2}} = \frac{\cos^2 \phi}{\rho} \\
\frac{\partial^2 \phi}{\partial y^2} &= -\frac{2xy}{\left(x^2+y^2\right)^2} = -\frac{2\sin\phi\cos\phi}{\rho^2}
\end{aligned}
$$

Finish the derivation by substituting these derivatives into the above expression.

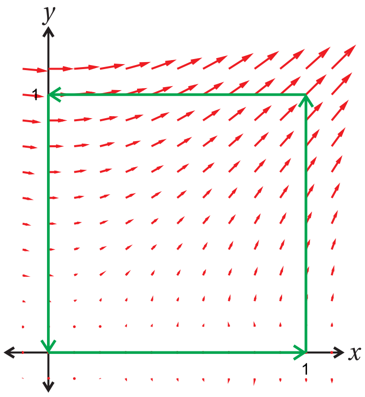


Figure 0.7

P0.11 Verify Stokes' theorem (0.12) for the function given in Example 0.3. Take the surface to be a square in the xy-plane contained by $x = 0$, $x = 1$, $y = 0$, and $y = 1$, as illustrated in Fig. 0.7.

P0.12 Verify the following vector integral theorem for the same volume used in Example 0.3, but with $\mathbf{F} = y^2 x\hat{\mathbf{x}} + xy\hat{\mathbf{z}}$ and $\mathbf{G} = x^2\hat{\mathbf{x}}$:

$$
\int_V \left[\mathbf{F}\left(\nabla\cdot\mathbf{G}\right) + \left(\mathbf{G}\cdot\nabla\right)\mathbf{F}\right] dv = \oint_S \mathbf{F}\left(\mathbf{G}\cdot\hat{\mathbf{n}}\right) da
$$

P0.13 Use the divergence theorem to show that the function in P0.5 is 4π times the three-dimensional delta function

$$
\delta^3\left(\mathbf{r}' - \mathbf{r}\right) \equiv \delta\left(x' - x\right)\delta\left(y' - y\right)\delta\left(z' - z\right)
$$

which has the property that

$$
\int_V \delta^3\left(\mathbf{r}' - \mathbf{r}\right) dv = \begin{cases} 1 \text{ if } V \text{ contains } \mathbf{r}' \\ 0 \text{ otherwise} \end{cases}
$$

Solution: We have by the divergence theorem

$$
\oint_S \frac{\left(\mathbf{r} - \mathbf{r}'\right)}{\left|\mathbf{r} - \mathbf{r}'\right|^3} \cdot \hat{\mathbf{n}} da = \int_V \nabla_{\mathbf{r}} \cdot \frac{\left(\mathbf{r} - \mathbf{r}'\right)}{\left|\mathbf{r} - \mathbf{r}'\right|^3} dv
$$

From P0.5, the argument in the integral on the right-hand side is zero except at $\mathbf{r} = \mathbf{r}'$. Therefore, if the volume V does not contain the point $\mathbf{r} = \mathbf{r}'$, then the result of both integrals must be zero. Let us construct a volume between an arbitrary surface S_1 containing $\mathbf{r} = \mathbf{r}'$and S_2, the surface of a tiny sphere centered on $\mathbf{r} = \mathbf{r}'$. Since the point $\mathbf{r} = \mathbf{r}'$ is excluded by the tiny sphere, the result of either integral in the divergence theorem is still zero. However, we have on the tiny sphere

$$\oint_{S_2} \frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3} \cdot \hat{\mathbf{n}} da = -\int_0^{2\pi}\int_0^{\pi} \left(\frac{1}{r_\epsilon^2}\right) r_\epsilon^2 \sin\phi \, d\phi \, d\alpha = -4\pi$$

Therefore, for the outer surface S_1 (containing $\mathbf{r} = \mathbf{r}'$) we must have the equal and opposite result:

$$\oint_{S_1} \frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3} \cdot \hat{\mathbf{n}} da = 4\pi$$

This implies

$$\int_V \nabla_{\mathbf{r}} \cdot \frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3} dv = \begin{cases} 4\pi \text{ if } V \text{ contains } \mathbf{r}' \\ 0 \text{ otherwise} \end{cases}$$

The integrand exhibits the same characteristics as the delta function Therefore, $\nabla_{\mathbf{r}} \cdot \frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3} = 4\pi\delta^3(\mathbf{r}-\mathbf{r}')$. The delta function is defined in (0.52)

Exercises for 0.2 Complex Numbers

P0.14 Let $z_1 = 1 - i$ and $z_2 = 3 + 4i$. (a) Compute $z_1 - z_2$ in rectangular form and convert the answer to polar form.

(b) Compute z_1/z_2 in rectangular form and convert the answer to polar form.

P0.15 Show that

$$\frac{a - ib}{a + ib} = e^{-2i\tan^{-1}\frac{b}{a}}$$

regardless of the sign of a, assuming a and b are real.

P0.16 Invert (0.15) to get both formulas in (0.16). HINT: You can get a second equation by considering Euler's equation with a negative angle $-\phi$.

P0.17 Show $\text{Re}\{A\} \times \text{Re}\{B\} = (AB + A^*B)/4 + C.C.$

P0.18 If $E_0 = |E_0|\, e^{i\delta_E}$ and $B_0 = |B_0|\, e^{i\delta_B}$, and if k, z, ω, and t are all real, prove

$$\text{Re}\left\{E_0 e^{i(kz-\omega t)}\right\}\text{Re}\left\{B_0 e^{i(kz-\omega t)}\right\} = \frac{1}{4}\left(E_0^* B_0 + E_0 B_0^*\right) + \frac{1}{2}|E_0|\,|B_0|\cos\left[2(kz-\omega t) + \delta_E + \delta_B\right]$$

P0.19 (a) If $\sin\phi = 2$, show that $\cos\phi = i\sqrt{3}$. HINT: Use $\sin^2\phi + \cos^2\phi = 1$.

(b) Show that the angle ϕ in (a) is $\pi/2 - i\ln(2+\sqrt{3})$.

P0.20 Write $A\cos(\omega t) + 2A\sin(\omega t + \pi/4)$ as a single phase-shifted cosine wave (i.e. find the amplitude and phase of the resultant cosine wave).

Exercises for 0.4 Fourier Theory

P0.21 Prove that Fourier Transforms have the property of linear superposition:

$$\mathscr{F}\left\{ag(t) + bh(t)\right\} = ag(\omega) + bh(\omega)$$

where $g(\omega) \equiv \mathscr{F}\left\{g(t)\right\}$ and $h(\omega) \equiv \mathscr{F}\{h(t)\}$.

P0.22 Prove $\mathscr{F}\left\{g(at)\right\} = \frac{1}{|a|} g\left(\frac{\omega}{a}\right)$.

P0.23 Prove $\mathscr{F}\left\{g(t-\tau)\right\} = g(\omega)e^{i\omega\tau}$.

P0.24 Show that the Fourier transform of $E(t) = E_0 e^{-\frac{t^2}{2T^2}} \cos\omega_0 t$ is

$$E(\omega) = \frac{TE_0}{2}\left(e^{-\frac{(\omega+\omega_0)^2}{2/T^2}} + e^{-\frac{(\omega-\omega_0)^2}{2/T^2}}\right)$$

P0.25 Take the inverse Fourier transform of the result in P0.24. Check that it returns exactly the original function.

P0.26 The following operation is referred to as the *convolution* of the functions $g(t)$ and $h(t)$:

$$g(t) \otimes h(t)\big|_\tau \equiv \int_{-\infty}^{\infty} g(t)h(\tau - t)\, dt$$

A convolution measures the overlap of $g(t)$ and a reversed $h(t)$ as a function of the offset τ. The result is a function of τ.

(a) Prove the convolution theorem:

$$\mathscr{F}\left\{g(t) \otimes h(t)\big|_\tau\right\}\Big|_\omega = \sqrt{2\pi}\, g(\omega)h(\omega)$$

(b) Prove this related form of the convolution theorem:

$$\mathscr{F}\left\{g(t)h(t)\right\}\big|_\omega = \frac{1}{\sqrt{2\pi}}\, g(\omega') \otimes h(\omega')\big|_\omega$$

Solution: Part (a)

$$\begin{aligned}
\mathscr{F}\left\{\int_{-\infty}^{\infty} g(t)h(\tau - t)\, dt\right\}\Bigg|_\omega &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} g(t)\, h(\tau - t)\, dt\right\} e^{i\omega\tau} d\tau \qquad \text{(Let } \tau = t' + t\text{)}\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(t)\, h\left(t'\right) e^{i\omega\left(t'+t\right)} dt\, dt'\\
&= \sqrt{2\pi}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(t)\, e^{i\omega t} dt \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} h\left(t'\right) e^{i\omega t'} dt'\\
&= \sqrt{2\pi}\, g(\omega)\, h(\omega)
\end{aligned}$$

P0.27 The following operation is called an *autocorrelation* of the function $h(t)$:

$$\int_{-\infty}^{\infty} h(t)h^*(t-\tau)dt$$

This is similar to the convolution operation described in P0.26, where $h(t)$ is integrated against an offset (unreversed) version of itself—hence the prefix "auto." Prove the autocorrelation theorem:

$$\mathscr{F}\left\{\int_{-\infty}^{\infty} h(t)h^*(t-\tau)dt\right\} = \sqrt{2\pi}\,|h(\omega)|^2$$

P0.28 (a) Compute the Fourier transform of a Gaussian function, $g(t) = e^{-t^2/2T^2}$. Do the integral by hand using the table in Appendix 0.A.

(b) Compute the Fourier transform of a sine function, $h(t) = \sin\omega_0 t$. Do the integral without a computer using $\sin(x) = (e^{ix} - e^{-ix})/2i$, combined with the integral formula (0.54).

(c) Use your results from parts (a) and (b) together with the convolution theorem from P0.26(b) to evaluate the Fourier transform of $f(t) = e^{-t^2/2T^2}\sin\omega_0 t$. (The answer should be similar to P0.24).

(d) Plot $f(t)$ and the imaginary part of its Fourier transform for the parameters $\omega_0 = 1$ and $T = 8$.

Chapter 1

Electromagnetic Phenomena

In 1861, James Maxwell assembled the various known relationships of electricity and magnetism into a concise[1] set of equations:[2]

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \qquad \text{(Gauss' Law)} \tag{1.1}$$

$$\nabla \cdot \mathbf{B} = 0 \qquad \text{(Gauss' Law for magnetism)} \tag{1.2}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \text{(Faraday's Law)} \tag{1.3}$$

$$\nabla \times \frac{\mathbf{B}}{\mu_0} = \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \mathbf{J} \qquad \text{(Ampere's Law revised by Maxwell)} \tag{1.4}$$

Here **E** and **B** represent electric and magnetic fields, respectively. The charge density ρ describes the *charge per volume* distributed through space.[3] The current density **J** describes the *motion* of charge density (in units of ρ times velocity). The constant ϵ_0 is called the *permittivity,* and the constant μ_0 is called the *permeability*. Taken together, these are known as *Maxwell's equations.*

After introducing a key revision of Ampere's law, Maxwell realized that together these equations comprise a complete self-consistent theory of electromagnetic phenomena. Moreover, the equations imply the existence of electromagnetic waves, which travel at the speed of light. Since the speed of light had been measured before Maxwell's time, it was immediately apparent (as was already suspected) that light is a high-frequency manifestation of the same phenomena that govern the influence of currents and charges upon each other. Previously, optics had been considered a topic quite separate from electricity and magnetism. Once the connection was made, it became clear that Maxwell's equations form the theoretical foundations of optics, and this is where we begin our study of light.

---

[1] In Maxwell's original notation, this set of equations was hardly concise, written without the convenience of modern vector notation or ∇. His formulation wouldn't fit easily on a T-shirt!

[2] See J. D. Jackson, *Classical Electrodynamics*, 3rd ed., p. 1 (New York: John Wiley, 1999) or the back cover of D. J. Griffiths, *Introduction to Electrodynamics*, 3rd ed. (New Jersey: Prentice-Hall, 1999).

[3] In other parts of this book, we use ρ for the radius in cylindrical coordinates, not to be confused with charge density, which makes an appearance only in this chapter.

1.1 Gauss' Law

The force on a point charge q located at $\mathbf{r}$ exerted by another point charge q' located at $\mathbf{r}'$ is

$$\mathbf{F} = q\mathbf{E}(\mathbf{r}) \tag{1.5}$$

where

$$\mathbf{E}(\mathbf{r}) = \frac{q'}{4\pi\epsilon_0}\frac{\left(\mathbf{r}-\mathbf{r}'\right)}{|\mathbf{r}-\mathbf{r}'|^3} \tag{1.6}$$

This relationship is known as *Coulomb's law.* The force is directed along the vector $\mathbf{r}-\mathbf{r}'$, which points from charge q' to q as seen in Fig. 1.1. The length or *magnitude* of this vector is given by $\left|\mathbf{r}-\mathbf{r}'\right|$ (i.e. the distance between q' and q). The familiar inverse square law can be seen by noting that $\left(\mathbf{r}-\mathbf{r}'\right)/\left|\mathbf{r}-\mathbf{r}'\right|$ is a unit vector. We have written the force in terms of an *electric field* $\mathbf{E}(\mathbf{r})$, which is defined throughout space (regardless of whether a second charge q is actually present). The permittivity ϵ_0 amounts to a proportionality constant.

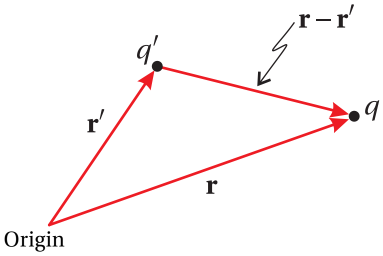


Figure 1.1 The geometry of Coulomb's law for a point charge.

The total force from a collection of charges is found by summing expression (1.5) over all charges q'_n associated with their specific locations $\mathbf{r}'_n$. If the charges are distributed continuously throughout space, having density $\rho\left(\mathbf{r}'\right)$ (units of charge per volume), the summation for finding the net electric field at $\mathbf{r}$ becomes an integral:

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\int_V \rho\left(\mathbf{r}'\right)\frac{\left(\mathbf{r}-\mathbf{r}'\right)}{|\mathbf{r}-\mathbf{r}'|^3}\,dv' \tag{1.7}$$

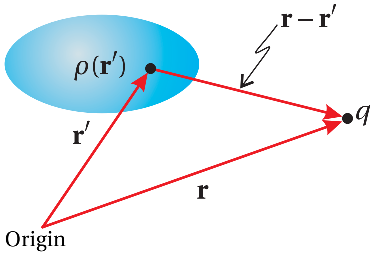


Figure 1.2 The geometry of Coulomb's law for a charge distribution.

This three-dimensional integral[4] gives the net electric field produced by the charge density ρ that exists in volume V.

Gauss' law (1.1), the first of Maxwell's equations, follows directly from (1.7) with some mathematical manipulation. No new physical phenomenon is introduced in this process.[5]

Derivation of Gauss' law

We begin with the divergence of (1.7):

$$\nabla\cdot\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\int_V \rho\left(\mathbf{r}'\right)\nabla_{\mathbf{r}}\cdot\frac{\left(\mathbf{r}-\mathbf{r}'\right)}{|\mathbf{r}-\mathbf{r}'|^3}\,dv' \tag{1.8}$$

The subscript on $\nabla_{\mathbf{r}}$ indicates that it operates on $\mathbf{r}$ while treating $\mathbf{r}'$, the dummy variable of integration, as a constant. The integrand contains a remarkable mathematical property that can be exploited, even without specifying the form of the charge distribution $\rho\left(\mathbf{r}'\right)$. In modern mathematical language, the vector expression in the integral is a three-dimensional delta function (see (0.52)):[6]

---

[4] Here dv' stands for $dx'dy'dz'$ and $\mathbf{r}' = x'\hat{\mathbf{x}} + y'\hat{\mathbf{y}} + z'\hat{\mathbf{z}}$ (in Cartesian coordinates).

[5] Actually, Coulomb's law applies only to static charge configurations, and in that sense it is incomplete since it implies an instantaneous response of the field to a reconfiguration of the charge. The generalized version of Coulomb's law, one of Jefimenko's equations, incorporates the fact that electromagnetic news travels at the speed of light. See D. J. Griffiths, *Introduction to Electrodynamics,* 3rd ed., Sect. 10.2.2 (New Jersey: Prentice-Hall, 1999). Ironically, Gauss' law, which can be derived from Coulomb's law, holds perfectly whether the charges remain still or are in motion.

$$\nabla_{\mathbf{r}} \cdot \frac{\left(\mathbf{r}-\mathbf{r}'\right)}{\left|\mathbf{r}-\mathbf{r}'\right|^3} \equiv 4\pi\delta^3\left(\mathbf{r}'-\mathbf{r}\right) \equiv 4\pi\delta\left(x'-x\right)\delta\left(y'-y\right)\delta\left(z'-z\right) \tag{1.9}$$

A derivation of this formula is addressed in problem P0.13. The delta function allows the integral in (1.8) to be performed, and the relation becomes simply

$$\nabla \cdot \mathbf{E}\left(\mathbf{r}\right) = \frac{\rho\left(\mathbf{r}\right)}{\epsilon_0}$$

which is the differential form of Gauss' law (1.1).

The (perhaps more familiar) integral form of Gauss' law can be obtained by integrating (1.1) over a volume V and applying the divergence theorem (0.11) to the left-hand side:

$$\oint_S \mathbf{E}\left(\mathbf{r}\right) \cdot \hat{\mathbf{n}}\, da = \frac{1}{\epsilon_0} \int_V \rho\left(\mathbf{r}\right)\, dv \tag{1.10}$$

This form of Gauss' law shows that the total electric field flux extruding through a closed surface S (i.e. the integral on the left side) is proportional to the net charge contained within it (i.e. within volume V contained by S).

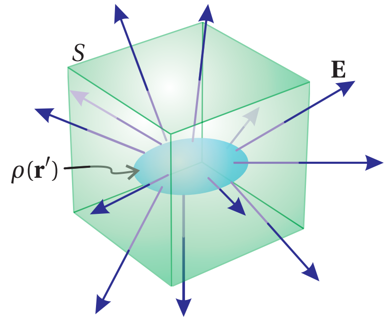


Figure 1.3 Gauss' law in integral form relates the flux of the electric field through a surface to the charge contained inside that surface.

Example 1.1

Suppose we have an electric field given by $\mathbf{E} = (\alpha x^2 y^3 \hat{\mathbf{x}} + \beta z^4 \hat{\mathbf{y}}) \cos\omega t$. Use Gauss' law (1.1) to find the charge density $\rho(x, y, z, t)$.

Solution:

$$\rho = \epsilon_0 \nabla \cdot \mathbf{E} = \epsilon_0 \left(\hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}\right)(\alpha x^2 y^3 \hat{\mathbf{x}} + \beta z^4 \hat{\mathbf{y}}) \cos\omega t = 2\epsilon_0 \alpha x y^3 \cos\omega t$$

1.2 Gauss' Law for Magnetic Fields

In order to 'feel' a magnetic force, a charge q must be moving at some velocity (call it $\mathbf{v}$). The *magnetic field* arises itself from charges that are in motion. We consider the magnetic field to arise from a distribution of moving charges described by a *current density* $\mathbf{J}\left(\mathbf{r}'\right)$ throughout space. The current density has units of charge times velocity per volume (or equivalently, current per cross sectional area). The magnetic force law analogous to Coulomb's law is

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B} \tag{1.11}$$

Carl Friedrich Gauss (1777–1855, German) was born in Braunschweig, Germany to a poor family. Gauss was a child prodigy, and he made his first significant advances to mathematics as a teenager. In grade school, he purportedly was asked to add all integers from 1 to 100, which he did in seconds to the astonishment of his teacher. (Presumably, Friedrich immediately realized that the numbers form fifty pairs equal to 101.) Gauss made important advances in number theory and differential geometry. He developed the law discussed here as one of Maxwell's equations in 1835, but it was not published until 1867, after Gauss' death. Ironically, Maxwell was already using Gauss' law by that time. (Wikipedia)

---

[6] For a derivation of Gauss' law from Coulomb's law that does not rely directly on the Dirac delta function, see J. D. Jackson, *Classical Electrodynamics* 3rd ed., pp. 27-29 (New York: John Wiley, 1999).

Jean-Baptiste Biot (1774-1862, French) was born in Paris. He attended the École Polytechnique where mathematician Gaspard Monge recognized his academic potential. After graduating, Biot joined the military and then took part in an insurrection on the side of the Royalists. He was captured, and his career might of have met a tragic ending there had Monge not successfully pleaded for his release from jail. Biot went on to become a professor of physics at the College de France. Among other contributions, Biot participated in the first hot-air balloon ride with Gay-Lussac and correctly deduced that meteorites that fell on L'Aigle, France in 1803 came from space. Later Biot collaborated with the younger Felix Savart (1791-1841) on the theory of magnetism and electrical currents. They formulated their famous law in 1820. (Wikipedia)

where

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int\limits_V \mathbf{J}(\mathbf{r}') \times \frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}\, dv' \tag{1.12}$$

The latter equation is known as the *Biot-Savart law*. The permeability μ_0 dictates the strength of the magnetic field, given the current distribution.

As with Coulomb's law, we can apply mathematics to the Biot-Savart law to obtain another of Maxwell's equations. Nevertheless, the essential physics is already inherent in the Biot-Savart law.[7] Using the result from P0.4 and P0.9(c), we can rewrite (1.12) as[8]

$$\mathbf{B}(\mathbf{r}) = -\frac{\mu_0}{4\pi}\int\limits_V \mathbf{J}(\mathbf{r}') \times \nabla_{\mathbf{r}}\frac{1}{|\mathbf{r}-\mathbf{r}'|}\, dv' = \frac{\mu_0}{4\pi}\nabla\times\int\limits_V \frac{\mathbf{J}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, dv' \tag{1.13}$$

Since the divergence of a curl is identically zero (see P0.6), we get straight away the second of Maxwell's equations (1.2)

$$\nabla\cdot\mathbf{B} = 0$$

which is known as Gauss' law for magnetic fields. (Two equations down; two to go.)

The similarity between $\nabla\cdot\mathbf{B} = 0$ and $\nabla\cdot\mathbf{E} = \rho/\epsilon_0$, Gauss' law for electric fields, is immediately apparent. In integral form, Gauss' law for magnetic fields looks the same as (1.10), only with zero on the right-hand side. If one were to imagine the existence of magnetic monopoles (i.e. isolated north or south 'charges'), then the right-hand side would not be zero. The law implies that the total magnetic flux extruding through any closed surface balances, with as many field lines pointing inwards as pointing outwards.

Example 1.2

The field surrounding a magnetic dipole is given by

$$\mathbf{B} = \beta\left[3xz\hat{\mathbf{x}} + 3yz\hat{\mathbf{y}} + \left(3z^2 - r^2\right)\hat{\mathbf{z}}\right]/r^5$$

where $r \equiv \sqrt{x^2+y^2+z^2}$. Show that this field satisfies Gauss' law for magnetic fields (1.2).

---

[7] Like Coulomb's law, the Biot-Savart law is incomplete since it also implies an instantaneous response of the magnetic field to a reconfiguration of the currents. The generalized version of the Biot-Savart law, another of Jefimenko's equations, incorporates the fact that electromagnetic news travels at the speed of light. Ironically, Gauss' law for magnetic fields and Maxwell's version of Ampere's law, derived from the Biot-Savart law, hold perfectly whether the currents are steady or vary in time. The Jefimenko equations, analogs of Coulomb and Biot-Savart, also embody Faraday's law, the only of Maxwell's equations that cannot be derived from the usual forms of Coulomb's law and the Biot-Savart law. See D. J. Griffiths, *Introduction to Electrodynamics,* 3rd ed., Sect. 10.2.2 (New Jersey: Prentice-Hall, 1999).

[8] Note that $\nabla_{\mathbf{r}}$ ignores the variable of integration $\mathbf{r}'$.

Solution:

$$\begin{aligned}
\nabla\cdot\mathbf{B} &= \beta\left[3\frac{\partial}{\partial x}\left(\frac{xz}{r^5}\right)+3\frac{\partial}{\partial y}\left(\frac{yz}{r^5}\right)+\frac{\partial}{\partial z}\left(\frac{3z^2}{r^5}-\frac{1}{r^3}\right)\right]\\
&= \beta\left[3\left(\frac{z}{r^5}-\frac{5xz}{r^6}\frac{\partial r}{\partial x}\right)+3\left(\frac{z}{r^5}-\frac{5yz}{r^6}\frac{\partial r}{\partial y}\right)+\left(\frac{6z}{r^5}-\frac{15z^2}{r^6}\frac{\partial r}{\partial z}+\frac{3}{r^4}\frac{\partial r}{\partial z}\right)\right]\\
&= \beta\left[\frac{12z}{r^5}-\frac{15z}{r^6}\left(x\frac{\partial r}{\partial x}+y\frac{\partial r}{\partial y}+z\frac{\partial r}{\partial z}\right)+\frac{3}{r^4}\frac{\partial r}{\partial z}\right]
\end{aligned}$$

The necessary derivatives are $\partial r/\partial x = x\Big/\sqrt{x^2+y^2+z^2} = x/r$, $\partial r/\partial y = y/r$, and $\partial r/\partial z = z/r$, which lead to

$$\nabla\cdot\mathbf{B} = \beta\left[\frac{12z}{r^5}-\frac{15z}{r^5}+\frac{3z}{r^5}\right]=0$$

1.3 Faraday's Law

Michael Faraday discovered that changing magnetic fields induce electric fields. This distinct physical effect, called induction, can be observed when a magnet is waved by a loop of wire. *Faraday's law* says that a change in magnetic flux through a circuit loop (see Fig. 1.4) induces a *voltage* around the loop according to

$$\oint_C \mathbf{E}\cdot d\ell = -\frac{\partial}{\partial t}\int_S \mathbf{B}\cdot\hat{\mathbf{n}}\,da \tag{1.14}$$

The right side describes a change in the magnetic flux through a surface, and the left side describes the voltage around the loop containing the surface.

We apply Stokes' theorem (0.12) to the left-hand side of Faraday's law and obtain

$$\int_S (\nabla\times\mathbf{E})\cdot\hat{\mathbf{n}}\,da = -\frac{\partial}{\partial t}\int_S \mathbf{B}\cdot\hat{\mathbf{n}}\,da \qquad \text{or} \qquad \int_S\left(\nabla\times\mathbf{E}+\frac{\partial\mathbf{B}}{\partial t}\right)\cdot\hat{\mathbf{n}}\,da=0 \tag{1.15}$$

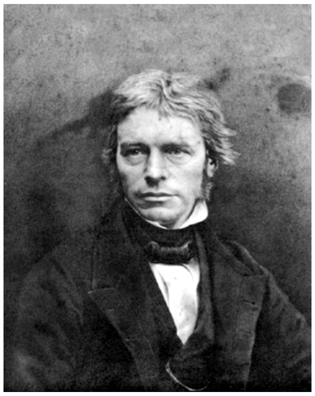
Since this equation is true regardless of what surface is chosen, it implies

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t}$$

which is the differential form of Faraday's law (1.4) (three of Maxwell's equations down; one to go).

Example 1.3

For the electric field given in Example 1.1, $\mathbf{E} = (\alpha x^2y^3\hat{\mathbf{x}}+\beta z^4\hat{\mathbf{y}})\cos\omega t$, use Faraday's law (1.3) to find $\mathbf{B}(x,y,z,t)$.



Michael Faraday (1791–1867, English) was one of the greatest experimental physicists in history. Born on the outskirts of London, his family was not well off, his father being a blacksmith. The young Michael Faraday only had access to a very basic education, and so he was mostly self-taught and never did acquire much skill in mathematics. As a teenager, he obtained a seven-year apprenticeship with a book binder, during which time he read many books, including books on science and electricity. Given his background, Faraday's entry into the scientific community was very gradual, from servant to assistant and eventually to director of the laboratory at the Royal Institution. Faraday is perhaps best known for his work that established the law of induction and for the discovery that magnetic fields can interact with light, known as the Faraday effect. He also made many advances to chemistry during his career including figuring out how to liquify several gases. Faraday was a deeply religious man, serving as a Deacon in his church. (Wikipedia)

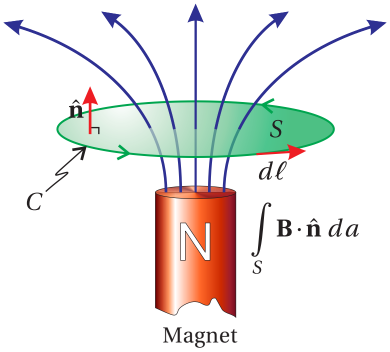


Figure 1.4 Faraday's law.

Solution:

$$\begin{aligned}\frac{\partial \mathbf{B}}{\partial t} &= -\nabla\times\mathbf{E} = -\cos\omega t\begin{vmatrix}\hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}}\\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z}\\ \alpha x^2y^3 & \beta z^4 & 0\end{vmatrix}\\ &= -\cos\omega t\left[\hat{\mathbf{x}}\frac{\partial}{\partial y}(0) - \hat{\mathbf{x}}\frac{\partial}{\partial z}\left(\beta z^4\right) - \hat{\mathbf{y}}\frac{\partial}{\partial x}(0) + \hat{\mathbf{y}}\frac{\partial}{\partial z}\left(\alpha x^2y^3\right)\right.\\ &\qquad\qquad\left. + \hat{\mathbf{z}}\frac{\partial}{\partial x}\left(\beta z^4\right) - \hat{\mathbf{z}}\frac{\partial}{\partial y}\left(\alpha x^2y^3\right)\right]\\ &= \left(4\beta z^3\hat{\mathbf{x}} + 3\alpha x^2y^2\hat{\mathbf{z}}\right)\cos\omega t\end{aligned}$$

Integrating in time, we get

$$\mathbf{B} = \left(4\beta z^3\hat{\mathbf{x}} + 3\alpha x^2y^2\hat{\mathbf{z}}\right)\frac{\sin\omega t}{\omega}$$

plus possibly a constant field.

André-Marie Ampère (1775-1836, French) was born in Lyon, France. The young André-Marie was tutored in Latin by his father, which gave him access to the mathematical works of Euler and Bernoulli to which he was drawn at an early age. When Ampère reached young adulthood, French revolutionaries executed his father. In 1799, Ampère married Julie Carron, who died of illness a few years later. These tragedies weighed heavy on Ampère throughout his life, especially because he was away from his wife during much of their short life together, while he worked as a professor of physics and chemistry in Bourg. After her death, Ampère was appointed professor of mathematics at the University of Lyon and then in 1809 at the Ècole Polytechnique in Paris. After hearing that a current-carrying wire could attract a compass needle in 1820, Ampère quickly developed the theory of electromagnetism. (Wikipedia)

1.4 Ampere's Law

The Biot-Savart law (1.12) can also be used to derive Ampere's law. Ampere's law is merely the inversion of the Biot-Savart law (1.12) so that **J** appears by itself, unfettered by integrals or the like.

Inversion of Biot-Savart Law

We take the curl of (1.12):

$$\nabla\times\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int_V \nabla_{\mathbf{r}}\times\left[\mathbf{J}\left(\mathbf{r}'\right)\times\frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}\right]dv' \tag{1.16}$$

We next apply the differential vector rule from P0.7 while noting that $\mathbf{J}\left(\mathbf{r}'\right)$ does not depend on $\mathbf{r}$ so that only two terms survive. The curl of $\mathbf{B}(\mathbf{r})$ then becomes

$$\nabla\times\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int_V\left(\mathbf{J}\left(\mathbf{r}'\right)\left[\nabla_{\mathbf{r}}\cdot\frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}\right] - \left[\mathbf{J}\left(\mathbf{r}'\right)\cdot\nabla_{\mathbf{r}}\right]\frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}\right)dv' \tag{1.17}$$

According to (1.9), the first term in the integral is $4\pi\mathbf{J}\left(\mathbf{r}'\right)\delta^3\left(\mathbf{r}'-\mathbf{r}\right)$, which is easily integrated. To make progress on the second term, we observe that the gradient can be changed to operate on the primed variables without affecting the final result (i.e. $\nabla_{\mathbf{r}}\to -\nabla_{\mathbf{r}'}$). In addition, we take advantage of a vector integral theorem (see P0.12) to arrive at

$$\nabla\times\mathbf{B}(\mathbf{r}) = \mu_0\mathbf{J}(\mathbf{r}) - \frac{\mu_0}{4\pi}\int_V\frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}\left[\nabla_{\mathbf{r}'}\cdot\mathbf{J}\left(\mathbf{r}'\right)\right]dv' + \frac{\mu_0}{4\pi}\oint_S\frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}\left[\mathbf{J}\left(\mathbf{r}'\right)\cdot\hat{\mathbf{n}}\right]da' \tag{1.18}$$

The last term in (1.18) vanishes if we assume that the current density **J** is completely contained within the volume V so that it is zero at the surface S. Thus, the expression for the curl of $\mathbf{B}(\mathbf{r})$ reduces to

$$\nabla \times \mathbf{B}(\mathbf{r}) = \mu_0 \mathbf{J}(\mathbf{r}) - \frac{\mu_0}{4\pi} \int\limits_V \frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3} \left[\nabla_{\mathbf{r}'} \cdot \mathbf{J}(\mathbf{r}')\right] dv' \tag{1.19}$$

As we will see in Section 1.5, $\nabla \cdot \mathbf{J} = 0$ for static charge distributions; in such cases the latter term in (1.19) vanishes

$$\nabla \cdot \mathbf{J} \cong 0 \qquad \text{(steady-state approximation)} \tag{1.20}$$

and we have succeeded in isolating **J** and obtained Ampere's law.

Without Maxwell's correction, Ampere's law

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} \tag{1.21}$$

only applies to quasi steady-state situations. The physical interpretation of Ampere's law is more apparent in integral form. We integrate both sides of (1.21) over an open surface S, bounded by contour C and apply Stokes' theorem (0.12) to the left-hand side:

$$\oint\limits_C \mathbf{B}(\mathbf{r}) \cdot d\boldsymbol{\ell} = \mu_0 \int\limits_S \mathbf{J}(\mathbf{r}) \cdot \hat{\mathbf{n}}\, da \equiv \mu_0 I \tag{1.22}$$

This law says that the line integral of **B** around a closed loop C is proportional to the total current flowing through the loop (see Fig. 1.5). The units of **J** are current per area, so the surface integral containing **J** yields the current I in units of charge per time.

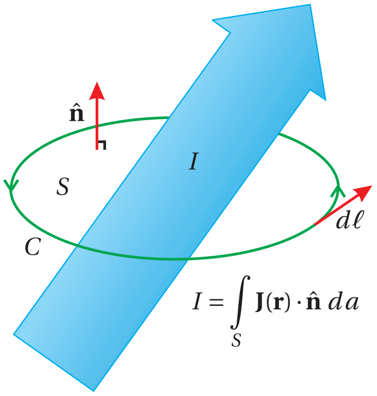

$\hat{\mathbf{n}}$
I
S
C
$d\ell$
$I = \int\limits_S \mathbf{J}(\mathbf{r}) \cdot \hat{\mathbf{n}}\, da$

Figure 1.5 Ampere's law.

1.5 Maxwell's Adjustment to Ampere's Law

Maxwell was the first to realize that Ampere's law was incomplete as written in (1.21) since there exist situations where $\nabla \cdot \mathbf{J} \neq 0$ (especially the case for optical phenomena). Maxwell figured out that (1.20) should be replaced with

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t} \tag{1.23}$$

This is called the *continuity equation* for charge and current densities. Simply stated, if there is net current flowing into a volume there ought to be charge piling up inside. For the steady-state situation inherently considered by Ampere, the current into and out of a volume is balanced so that $\partial \rho / \partial t = 0$.

Derivation of the Continuity Equation

Consider a volume of space enclosed by a surface S through which current is flowing. The total current exiting the volume is

$$I = \oint_S \mathbf{J} \cdot \hat{\mathbf{n}}\, da \tag{1.24}$$

where $\hat{\mathbf{n}}$ is the outward normal to the surface. The units on this equation are that of current, or charge per time, leaving the volume.

Since we have considered a *closed* surface S, the net current leaving the enclosed volume V must be the same as the rate at which charge within the volume vanishes:

$$I = -\frac{\partial}{\partial t} \int_V \rho\, dv \tag{1.25}$$

Upon equating these two expressions for current, as well as applying the divergence theorem (0.11) to the former, we get

$$\int_V \nabla \cdot \mathbf{J} dv = -\int_V \frac{\partial \rho}{\partial t} dv \qquad \text{or} \qquad \int_V \left(\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} \right) dv = 0 \tag{1.26}$$

Since (1.26) is true regardless of which volume V we choose, it implies (1.23).

James Clerk Maxwell (1831–1879, Scottish) was born to a wealthy family in Edinburgh, Scotland. Originally, his name was John Clerk, but he added his mother's maiden name when he inherited an estate from her family. Maxwell was a bright and inquisitive child and displayed an unusual gift for mathematics at an early age. He attended Edinburgh University and then Trinity College at Cambridge University. Maxwell started his career as a professor at Aberdeen University, but lost his job a few years later during restructuring, at which time Maxwell took a post at King's College of London. Maxwell is best known for his fundamental contributions to electricity and magnetism and the kinetic theory of gases. He studied numerous other subjects, including the human perception of color and color-blindness, and is credited with producing the first color photograph. He originally postulated that electromagnetic waves propagated in a mechanical "luminiferous ether." He founded the Cavendish laboratory at Cambridge in 1874, which has produced 28 Nobel prizes to date. Maxwell, one of Einstein's heros, died of stomach cancer in his forties. (Wikipedia)

Maxwell's main contribution (aside from organizing other people's formulas[9] and recognizing them as a complete set of coupled differential equations—a big deal) was the injection of the continuity equation (1.23) into the derivation of Ampere's law (1.19). This yields

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \frac{\mu_0}{4\pi} \frac{\partial}{\partial t} \int_V \rho\left(\mathbf{r}'\right) \frac{\left(\mathbf{r} - \mathbf{r}'\right)}{\left|\mathbf{r} - \mathbf{r}'\right|^3}\, dv' \tag{1.27}$$

Then substitution of (1.7) into this formula gives

$$\nabla \times \frac{\mathbf{B}}{\mu_0} = \mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}$$

the last of Maxwell's equations (1.4).

This revised version of Ampere's law includes the additional term $\epsilon_0 \partial \mathbf{E}/\partial t$, which is known as the *displacement current* (density). The displacement current exists even in the absence of any actual charge density ρ.[10] It indicates that a changing electric field behaves like a current in the sense that it produces magnetic fields. The similarity between Faraday's law and the corrected Ampere's law (1.4) is apparent. No doubt this played a part in motivating Maxwell's work.

---

[9] Although Gauss developed his law in 1835, it was not published until after his death in 1867, well after Maxwell published his laws of electromagnetism, so in practice Maxwell accomplished much more than merely fixing Ampere's law.

[10] Based on (1.27), one might think that the displacement current $\epsilon_0 \partial \mathbf{E}/\partial t$ ought to be zero in a region of space with no charge density ρ. However, in (1.27) ρ appears in a volume integral over a region of space sufficiently large (consistent with a previous supposition) to include any charges responsible for the field $\mathbf{E}$; presumably, all fields arise from sources.

In summary, in the previous section we saw that the basic physics in Ampere's law is present in the Biot-Savart law. Infusing it with charge conservation (1.23) yields the corrected form of Ampere's law.

Example 1.4

(a) Use Gauss' law to find the electric field in a gap that interrupts a current-carrying wire, as shown in Fig. 1.6.

(b) Find the strength of the magnetic field on contour C using Ampere's law applied to surface S_1.

(c) Show that the displacement current in the gap leads to the identical magnetic field when using surface S_2.

Solution: (a) We'll assume that the cross-sectional area of the wire A is much wider than the gap separation. Then the electric field in the gap will be uniform, and the integral on the left-hand side of (1.10) reduces to EA since there is essentially no field other than in the gap. If the accumulated charge on the 'plate' is Q, then the right-hand side of (1.10) integrates to Q/ϵ_0, and the electric field turns out to be $E = Q/(\epsilon_0 A)$.

(b) Let the contour C be a circle at radius r. The magnetic field points around the circumference with constant strength. The left-hand side of (1.22) becomes $2\pi r B$ while the right-hand side is

$$\mu_0 \int_S \mathbf{J}\cdot\hat{\mathbf{n}}\, da = \mu_0 I = \mu_0 \frac{\partial Q}{\partial t}$$

This gives for the magnetic field

$$B = \frac{\mu_0}{2\pi r}\frac{\partial Q}{\partial t}$$

(c) If instead we use the displacement current $\epsilon_0 \partial\mathbf{E}/\partial t$ in place of $\mathbf{J}$ in the right-hand side of right-hand side of (1.22), we get for that piece

$$\mu_0 \int_S \left(\epsilon_0 \frac{\partial \mathbf{E}}{\partial t}\right)\cdot\hat{\mathbf{n}}\, da = \mu_0\epsilon_0 \frac{\partial E}{\partial t} A = \mu_0 \frac{\partial Q}{\partial t}$$

which is the same as before.

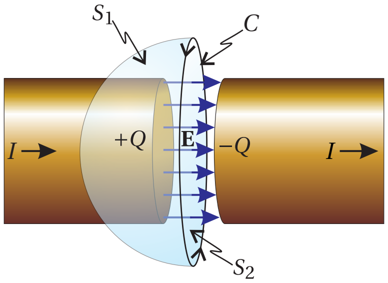


Figure 1.6 Charging capacitor.

Example 1.5

For the electric field $\mathbf{E} = (\alpha x^2 y^3 \hat{\mathbf{x}} + \beta z^4 \hat{\mathbf{y}})\cos\omega t$ (see Example 1.1) and the associated magnetic field $\mathbf{B} = \left(4\beta z^3 \hat{\mathbf{x}} + 3\alpha x^2 y^2 \hat{\mathbf{z}}\right)\frac{\sin\omega t}{\omega}$ (see Example 1.3), find the current density $J(x, y, z, t)$.

Solution:

$$\mathbf{J} = \nabla \times \frac{\vec{B}}{\mu_0} - \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \frac{\sin \omega t}{\mu_0 \omega} \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ 4\beta z^3 & 0 & 3\alpha x^2 y^2 \end{vmatrix} + \epsilon_0 \omega (\alpha x^2 y^3 \hat{\mathbf{x}} + \beta z^4 \hat{\mathbf{y}}) \sin \omega t$$

$$= \frac{\sin \omega t}{\mu_0 \omega} \left[6\alpha x^2 y \hat{\mathbf{x}} - 6\alpha x y^2 \hat{\mathbf{y}} + 12\beta z^2 \hat{\mathbf{y}}\right] + \epsilon_0 \omega (\alpha x^2 y^3 \hat{\mathbf{x}} + \beta z^4 \hat{\mathbf{y}}) \sin \omega t$$

$$= \left[\left(\epsilon_0 \omega \alpha x^2 y^3 + \frac{6\alpha x^2 y}{\mu_0 \omega}\right)\hat{\mathbf{x}} + \left(\epsilon_0 \omega \beta z^4 + \frac{12\beta z^2}{\mu_0 \omega} - \frac{6\alpha x y^2}{\mu_0 \omega}\right)\hat{\mathbf{y}}\right] \sin \omega t$$

1.6 Polarization of Materials

We are essentially finished with our analysis of Maxwell's equations except for a brief discussion of current density **J** and charge density ρ. For convenience, it is common to decompose the current density into three categories:

$$\mathbf{J} = \mathbf{J}_{\text{free}} + \mathbf{J}_{\text{m}} + \mathbf{J}_{\text{p}} \tag{1.28}$$

First, as you might expect, currents can arise from free charges in motion such as electrons in a metal, referred to as $\mathbf{J}_{\text{free}}$. Second, individual atoms can exhibit internal currents that give rise to paramagnetic and diamagnetic effects, denoted by $\mathbf{J}_{\text{m}}$. These are seldom important in optics problems, and so we will ignore these types of currents by writing $\mathbf{J}_{\text{m}} = 0$. Third, molecules in a material can elongate, becoming dipoles in response to an applied electric field. This gives rise to a *polarization current* $\mathbf{J}_{\text{p}}$.

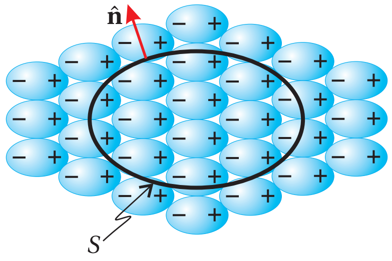


Figure 1.7 A polarized medium with $\nabla \cdot \mathbf{P} = 0$.

The polarization current is associated with a dipole distribution function **P**, called the *polarization*[11] (in units of dipoles per volume, or charge times length per volume). Physically, if the dipoles (depicted in Fig. 1.7) change their strength or orientation as a function of time, an effective current density arises in the medium. Note that the time-derivative of an individual dipole moment renders charge times velocity. Thus, the time derivative of 'sloshing' dipoles per volume gives a current density equal to

$$\mathbf{J}_{\text{p}} = \frac{\partial \mathbf{P}}{\partial t} \tag{1.29}$$

Next, we turn our attention to the charge density, which is often decomposed as

$$\rho = \rho_{\text{free}} + \rho_{\text{p}} \tag{1.30}$$

We seldom consider the propagation of electromagnetic waveforms through electrically charged materials, and so in this book we will always write $\rho_{\text{free}} = 0$. One might be tempted in this case to assume that the overall charge density is zero, but this would be wrong. Even though a material is electrically neutral, the polarization **P** can vary in space, leading to local concentrations of positive or negative charges. This type of charge density is denoted by ρ_p. It arises from nonuniform arrangements of dipoles, as depicted in Fig. 1.8.

---

[11] Unfortunately, the word *polarization* gets double usage. It also refers to the orientation of the electric field in electromagnetic waves, which is the topic of chapter 6.

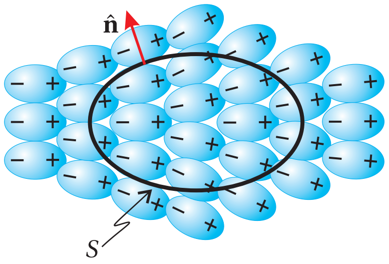


Figure 1.8 A polarized medium with $\nabla \cdot \mathbf{P} \neq 0$.

To connect ρ_p with **P**, we write the continuity equation (1.23) for the current and charge densities associated with the polarization:

$$\nabla \cdot \mathbf{J}_\mathrm{p} = -\frac{\partial \rho_\mathrm{p}}{\partial t} \tag{1.31}$$

Substitution of (1.29) into this equation immediately yields

$$\rho_\mathrm{p} = -\nabla \cdot \mathbf{P} \tag{1.32}$$

Example 1.6 To better appreciate local buildup of charge due to variation in the medium polarization, consider the divergence theorem (0.11) applied to **P**:

$$-\oint_S \mathbf{P}(\mathbf{r}) \cdot \hat{\mathbf{n}}\, da = -\int_V \nabla \cdot \mathbf{P}(\mathbf{r})\, dv$$

The left-hand side is a surface integral, which after integrating gives units of charge. Physically, it is the sum of the charges touching the inside of surface S (multiplied by a minus since by convention dipole vectors point from the negatively charged end of a molecule to the positively charged end). When $\nabla \cdot \mathbf{P}$ is zero, there are equal numbers of positive and negative charges touching S from within, as depicted in Fig. 1.7. When $\nabla \cdot \mathbf{P}$ is not zero, the positive and negative charges touching S are not balanced, as depicted in Fig. 1.8. Essentially, excess charge ends up within the volume because the nonuniform alignment of dipoles causes them to be cut preferentially at the surface.[12]

In summary, in electrically neutral nonmagnetic media, Maxwell's equations (in terms of the medium polarization **P**) are[13]

$$\nabla \cdot \mathbf{E} = -\frac{\nabla \cdot \mathbf{P}}{\epsilon_0} \qquad \text{(Gauss' law)} \tag{1.33}$$

---

[12] The figures may give the impression that you could always just draw a surface that avoids cutting any dipoles. However, the function $\mathbf{P}(\mathbf{r})$ is continuous, while the figures depict crudely just a few dipoles. In a continuous material you can't draw a surface that avoids cutting dipoles.

[13] It is not uncommon to see the macroscopic Maxwell equations written in terms of two auxiliary fields: **H** and **D**. The field **H** is useful in magnetic materials. In these materials, the combination $\mathbf{B}/\mu_0$ in Ampere's law is replaced by $\mathbf{H} \equiv \mathbf{B}/\mu_0 - \mathbf{M}$, where $\mathbf{J}_\mathrm{m} = \nabla \times \mathbf{M}$ is the current associated with the material's magnetization. Since we only consider nonmagnetic materials ($\mathbf{M} = 0$), there is little point in using **H**. The field **D**, called the displacement, is defined as $\mathbf{D} \equiv \epsilon_0 \mathbf{E} + \mathbf{P}$. This combination of **E** and **P** occurs in Coulomb's law and Ampere's law. For physical clarity, the authors of this book elect to retain the prominence of the polarization **P** in the equations.

$$\nabla \cdot \mathbf{B} = 0 \qquad \text{(Gauss' law for magnetism)} \tag{1.34}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \text{(Faraday's law)} \tag{1.35}$$

$$\nabla \times \frac{\mathbf{B}}{\mu_0} = \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}}{\partial t} + \mathbf{J}_{\text{free}} \qquad \text{(Ampere's law; fixed by Maxwell)} \tag{1.36}$$

1.7 The Wave Equation

When Maxwell unified electromagnetic theory, he immediately noticed that waves are solutions to his set of equations. In fact, his desire to find a set of equations that allowed for waves aided his effort to find the correct equations. After all, it was already known that light traveled as waves. Kirchhoff had previously pointed out that $1\big/\sqrt{\epsilon_0\mu_0}$ gives the correct speed of light $c = 3 \times 10^8$ m/s (which had previously been measured). Faraday and Kerr had observed that strong magnetic and electric fields affect light propagating in crystals. The time was right to suspect that light was an electromagnetic phenomena taking place at high frequency.

At first glance, Maxwell's equations might not immediately suggest (to the inexperienced eye) that waves are solutions. However, we can manipulate the equations (first order differential equations that couple **E** to **B**) into the familiar wave equation (decoupled second order differential equations for either **E** or **B**). You should become familiar with this derivation. In what follows, we will derive the wave equation for **E**. The derivation of the wave equation for **B** is very similar (see problem P1.6).

Derivation of the Wave Equation

Taking the curl of (1.3) gives

$$\nabla \times (\nabla \times \mathbf{E}) + \frac{\partial}{\partial t}(\nabla \times \mathbf{B}) = 0 \tag{1.37}$$

We may eliminate $\nabla \times \mathbf{B}$ by substitution from (1.4), which gives

$$\nabla \times (\nabla \times \mathbf{E}) + \mu_0\epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = -\mu_0 \frac{\partial \mathbf{J}}{\partial t} \tag{1.38}$$

Next we apply the vector identity (0.10), $\nabla \times (\nabla \times \mathbf{E}) = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E}$, and use Gauss' law (1.1) to replace the term $\nabla \cdot \mathbf{E}$, which brings us to

$$\nabla^2 \mathbf{E} - \mu_0\epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = \mu_0 \frac{\partial \mathbf{J}}{\partial t} + \frac{\nabla \rho}{\epsilon_0} \tag{1.39}$$

If we perform the above derivation starting from (1.33)–(1.36) (or equivalently, if we substitute (1.28)–(1.32) into (1.39)), we obtain a form that is more useful for optics:

$$\nabla^2 \mathbf{E} - \mu_0\epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = \mu_0 \frac{\partial \mathbf{J}_{\text{free}}}{\partial t} + \mu_0 \frac{\partial^2 \mathbf{P}}{\partial t^2} - \frac{1}{\epsilon_0}\nabla(\nabla \cdot \mathbf{P}) \tag{1.40}$$

The left-hand side of (1.40), when set to zero, is the familiar wave equation. However, the right-hand side contains a number of 'source terms', which arise when various currents and/or polarizations are present. The first term on the right-hand side of (1.40) describes currents of free charges, which are important for determining the reflection of light from a metallic surface or for determining the propagation of light in a plasma. The second term on the right-hand side describes dipole oscillations, which behave similar to currents. These dipole oscillations play a prominent role when light propagates in nonconducting materials. The final term on the right-hand side of (1.40) is important in anisotropic media such as crystals. In this case, the polarization **P** responds to the electric field along a direction not necessarily parallel to **E**, due to the influence of the crystal lattice (addressed in chapter 5).

In summary, when light propagates in a material, at least one of the terms on the right-hand side of (1.40) will be nonzero. As an example, in glass, $\mathbf{J}_{\text{free}} = 0$ and $\nabla \cdot \mathbf{P} = 0$, but $\partial^2 \mathbf{P} / \partial t^2 \neq 0$ since the medium polarization responds to the light field, giving rise to refractive index (discussed in chapter 2).

Example 1.7

Show that the electric field

$$\mathbf{E} = (\alpha x^2 y^3 \hat{\mathbf{x}} + \beta z^4 \hat{\mathbf{y}}) \cos \omega t$$

and the associated charge density (see Example 1.1)

$$\rho = 2\epsilon_0 \alpha x y^3 \cos \omega t$$

together with the associated current density (see Example 1.5)

$$\mathbf{J} = \left[\left(\epsilon_0 \omega \alpha x^2 y^3 + \frac{6\alpha x^2 y}{\mu_0 \omega} \right) \hat{\mathbf{x}} + \left(\epsilon_0 \omega \beta z^4 + \frac{12\beta z^2}{\mu_0 \omega} - \frac{6\alpha x y^2}{\mu_0 \omega} \right) \hat{\mathbf{y}} \right] \sin \omega t$$

satisfy the wave equation (1.39).

Solution: We have

$$\begin{aligned} \nabla^2 \mathbf{E} - \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} &= \left[\alpha \left(2y^3 + 6x^2 y \right) \hat{\mathbf{x}} + 12\beta z^2 \hat{\mathbf{y}} \right] \cos \omega t \\ &\quad + \mu_0 \epsilon_0 \omega^2 (\alpha x^2 y^3 \hat{\mathbf{x}} + \beta z^4 \hat{\mathbf{y}}) \cos \omega t \\ &= \left[\alpha \left(2y^3 + 6x^2 y + \mu_0 \epsilon_0 \omega^2 x^2 y^3 \right) \hat{\mathbf{x}} + \beta \left(12z^2 + \mu_0 \epsilon_0 \omega^2 z^4 \right) \hat{\mathbf{y}} \right] \cos \omega t \end{aligned}$$

Similarly,

$$\begin{aligned} \mu_0 \frac{\partial \mathbf{J}}{\partial t} + \frac{\nabla \rho}{\epsilon_0} &= \left[\left(\mu_0 \epsilon_0 \omega^2 \alpha x^2 y^3 + 6\alpha x^2 y \right) \hat{\mathbf{x}} + \left(\mu_0 \epsilon_0 \omega^2 \beta z^4 + 12\beta z^2 - 6\alpha x y^2 \right) \hat{\mathbf{y}} \right] \cos \omega t \\ &\quad + \left[2\alpha y^3 \hat{\mathbf{x}} + 6\alpha x y^2 \hat{\mathbf{y}} \right] \cos \omega t \\ &= \left[\alpha \left(2y^3 + 6x^2 y + \mu_0 \epsilon_0 \omega^2 x^2 y^3 \right) \hat{\mathbf{x}} + \beta \left(12z^2 + \mu_0 \epsilon_0 \omega^2 z^4 \right) \hat{\mathbf{y}} \right] \cos \omega t \end{aligned}$$

The two expressions are identical, and the wave equation is satisfied.[14]

---

[14] The expressions in Example 1.7 hardly look like waves. The (quite unlikely) current and charge distributions, which fill all space, would have to be artificially induced rather than arise naturally in response to a field disturbance on a medium.

The magnetic field **B** satisfies a similar wave equation, decoupled from **E** (see P1.6). However, the two waves are not independent. The fields for **E** and **B** must be chosen to be consistent with each other through Maxwell's equations. After solving the wave equation (1.40) for **E**, one can obtain the consistent **B** from **E** via Faraday's law (1.35).

In vacuum, all of the terms on the right-hand side in (1.40) are zero. In this case, the wave equation reduces to

$$\nabla^2 \mathbf{E} - \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0 \qquad \text{(vacuum)} \quad (1.41)$$

Solutions to this equation can take on every imaginable functional shape (specified at a given instant—the evolution thereafter being controlled by (1.41)). Moreover, since the differential equation is linear, any number of solutions can be added together to create other valid solutions. Consider the subclass of solutions that propagate in a particular direction. These waveforms preserve shape while traveling with speed

$$c \equiv 1 \big/ \sqrt{\epsilon_0 \mu_0} = 2.9979 \times 10^8 \text{ m/s} \qquad (1.42)$$

In this case, **E** depends on the argument $\hat{\mathbf{u}} \cdot \mathbf{r} - ct$, where $\hat{\mathbf{u}}$ is a unit vector specifying the direction of propagation. The shape is preserved since features occurring at a given position recur 'downstream' at a distance ct after a time t. By checking this solution in (1.41), one confirms that the speed of propagation is c (see P1.8). As mentioned previously, one may add together any combination of solutions (even with differing directions of propagation) to form other valid solutions.

Exercises

Exercises for 1.1 Gauss' Law

P1.1 Consider an infinitely long hollow cylinder with inner radius a and outer radius b as shown in Fig. 1.9. Assume that the cylinder has a charge density $\rho = k/s^2$ for $a < s < b$ and no charge elsewhere, where s is the radial distance from the axis of the cylinder. Use Gauss' Law in integral form to find the electric field produced by this charge for each of the three regions: $s < a$, $a < s < b$, and $s > b$.

HINT: For each region first draw an appropriate 'Gaussian surface' and integrate the charge density over the volume to figure out the enclosed charge. Then use Gauss' law in integral form and the symmetry of the problem to solve for the electric field.

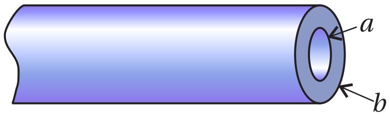


Figure 1.9 A charged cylinder with charge located between a and b.

Exercises for 1.3 Faraday's Law

P1.2 Suppose that an electric field is given by $\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0 \cos\left(\mathbf{k}\cdot\mathbf{r} - \omega t + \phi\right)$, where $\mathbf{k}\perp\mathbf{E}_0$ and ϕ is a constant phase. Show that

$$\mathbf{B}(\mathbf{r},t) = \frac{\mathbf{k}\times\mathbf{E}_0}{\omega}\cos\left(\mathbf{k}\cdot\mathbf{r} - \omega t + \phi\right)$$

is consistent with (1.3).

Exercises for 1.4 Ampere's Law

P1.3 A conducting cylinder with the same geometry as P1.1 carries a current density $\mathbf{J} = k/s\hat{\mathbf{z}}$ along the axis of the cylinder for $a < s < b$, where s is the radial distance from the axis of the cylinder. Using Ampere's Law in integral form, find the magnetic field due to this current in regions (a) $s < a$, (b) $a < s < b$, and (c) $s > b$.

HINT: For each region first draw an appropriate 'Amperian loop' and integrate the current density over the surface to figure out how much current passes through the loop. Then use Ampere's law in integral form and the symmetry of the problem to solve for the magnetic field.

Exercises for 1.6 Polarization of Materials

P1.4 Check that the **E** and **B** fields in P1.2 satisfy the rest of Maxwell's equations:

(a) (1.1). What must ρ be?

(b) (1.2).

(c) (1.4). What must **J** be?

P1.5 Memorize Maxwell's equations (1.33)–(1.36). Be prepared to reproduce them from memory on an exam, and write them on your homework from memory to indicate completion. Also very briefly summarize the physical principles described by each of Maxwell's equations, and the approximations made to (1.28) and (1.30).

Exercises for 1.7 The Wave Equation

P1.6 Derive the wave equation for the magnetic field $\mathbf{B}$ in vacuum (i.e. $\mathbf{J} = 0$ and $\rho = 0$).

P1.7 Show that the magnetic field in P1.2 is consistent with the wave equation derived in P1.6. What is the requirement on k and ω?

P1.8 Verify that $\mathbf{E}(\hat{\mathbf{u}} \cdot \mathbf{r} - ct)$ satisfies the vacuum wave equation (1.41), where $\mathbf{E}$ has an *arbitrary* functional form.

P1.9 (a) Show that $\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 \cos\left(k(\hat{\mathbf{u}} \cdot \mathbf{r} - ct) + \phi\right)$ is a solution to the vacuum wave equation (1.41), where $\hat{\mathbf{u}}$ is an arbitrary unit vector, $c = 1\big/\sqrt{\epsilon_0 \mu_0}$, and k is a constant with units of inverse length.

(b) Show that each wavefront forms a plane, which is why such solutions are often called 'plane waves'. HINT: A wavefront is a surface in space where the argument of the cosine (i.e. the *phase* of the wave) has a constant value. Set the cosine argument to an arbitrary constant and see what positions are associated with that phase.

(c) Determine the speed $v = \Delta r / \Delta t$ that a wavefront moves in the $\hat{\mathbf{u}}$ direction. HINT: Set the cosine argument to a constant, and consider a change in position along $\hat{\mathbf{u}}$ with its associated change in time.

(d) By analysis of this wave, determine the wavelength λ in terms of k. HINT: Holding time constant, find the distance between identical wavefronts by changing the position along $\hat{\mathbf{u}}$ and allowing the cosine argument to evolve through 2π.

(e) Use (1.33) to show that $\mathbf{E}_0$ and $\hat{\mathbf{u}}$ must be perpendicular to each other in vacuum.

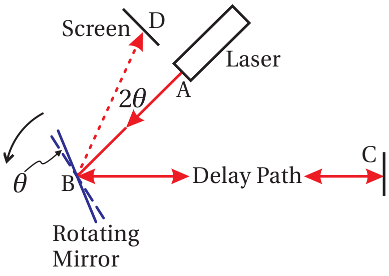


Figure 1.10 Geometry for lab 1.10.

L1.10 Measure the speed of light using a rotating mirror. Provide an estimate of the experimental uncertainty in your answer (not the percentage error from the known value). (video)

Figure 1.10 shows a simplified geometry for the optical path for light in this experiment. Laser light from A reflects from a rotating mirror at B towards C. The light returns to B, where the mirror has rotated, sending the light to point D. Notice that a mirror rotation of θ deflects the beam by 2θ.

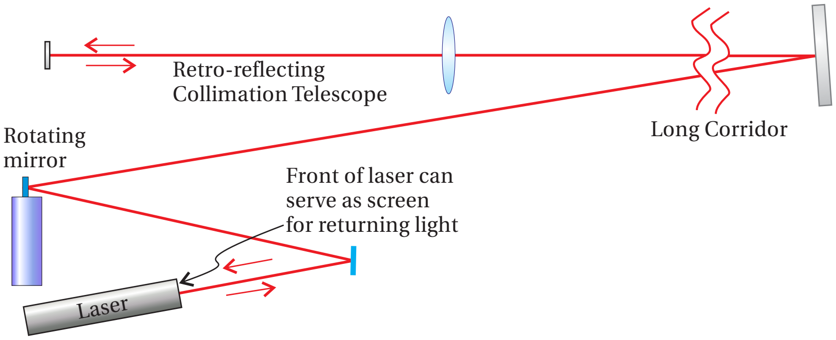


Figure 1.11 A schematic of the setup for lab 1.10.

Ole Roemer (1644–1710, Danish) was a man of many interests. In addition to measuring the speed of light, he created a temperature scale which with slight modification became the Fahrenheit scale, introduced a system of standard weights and measures, and was heavily involved in civic affairs (city planning, etc.). Scientists initially became interested in Io's orbit because its eclipse (when it went behind Jupiter) was an event that could be seen from many places on earth. By comparing accurate measurements of the local time when Io was eclipsed by Jupiter at two remote places on earth, scientists in the 1600s were able to determine the longitude difference between the two places.

P1.11 Ole Roemer made the first successful measurement of the speed of light in 1676 by observing the orbital period of Io, a moon of Jupiter with a period of 42.5 hours. When Earth is moving toward Jupiter, the period is measured to be slightly less, owing to decreasing Jupiter-Earth distance between successive Io orbits. When Earth is moving away from Jupiter, the situation is reversed, and the period is measured to be slightly longer.

(a) If you were to measure the time for 40 observed orbits of Io when Earth is moving directly toward Jupiter and then later measure the time for 40 observed orbits when Earth is moving directly away from Jupiter, what would you expect the difference between these two measurements to be? Take the Earth's orbital radius to be 1.5×10^{11} m. To simplify the geometry, just assume that Earth moves directly toward or away from Jupiter over the entire 40 orbits (see Fig. 1.12).

(b) Roemer did the experiment described in part (a), and experimentally measured a 22 minute difference. What speed of light would one deduce from that value?

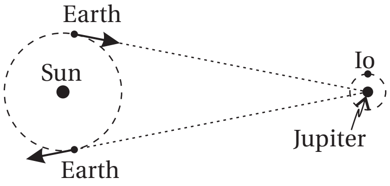


Figure 1.12 Geometry for P1.11

P1.12 In an isotropic nonconducting medium (i.e. $\nabla\cdot\mathbf{P}=0$, $\mathbf{J}_{\text{free}}=0$), the polarization under certain assumptions can be written as function of the electric field: $\mathbf{P}=\epsilon_0\chi(E)\,\mathbf{E}$, where $\chi(E)=\chi_1+\chi_2E+\chi_3E^2\cdots$. The higher order coefficients in the expansion (i.e. χ_2, χ_3, ...) are typically small, so only the first term is important unless the field is very strong. *Nonlinear optics* deals with the study of intense light-matter interactions, where the higher-order terms in the expansion become important. This can lead to phenomena such as harmonic generation.

Starting with (1.40), show that the wave equation becomes:

$$\nabla^2\mathbf{E}-\mu_0\epsilon_0\left(1+\chi_1\right)\frac{\partial^2\mathbf{E}}{\partial t^2}=\mu_0\epsilon_0\frac{\partial^2\left(\chi_2E+\chi_3E^2+\cdots\right)\mathbf{E}}{\partial t^2}$$

Chapter 2

Plane Waves and Refractive Index

In this chapter, we study sinusoidal solutions of Maxwell's equations, called *plane waves*. Restricting our attention to plane waves may seem limiting at first, since (as mentioned in chapter 1) an endless variety of waveform shapes can satisfy the wave equation in vacuum. It turns out, however, that an arbitrary waveform can always be constructed from a linear superposition of sinusoidal waves. Thus, there is no loss of generality if we focus our attention on plane-wave solutions.

In a material, the electric field of a plane wave induces oscillating dipoles, and these oscillating dipoles in turn alter the electric field. We use the *index of refraction* to describe this effect. Plane waves of different frequencies experience different refractive indices, which causes them to travel at different speeds in materials. Thus, an arbitrary waveform, which is composed of multiple sinusoidal waves, invariably changes shape as it travels in a material, as the different sinusoidal waves change relationship with respect to one another. This *dispersion* phenomenon is a primary reason why physicists and engineers choose to work with sinusoidal waves. Every waveform except for individual sinusoidal waves changes shape as it travels in a material.

When describing plane waves, it is convenient to employ complex numbers to represent physical quantities. This is particularly true for problems involving absorption, which takes place in metals and, to a lesser degree (usually), in *dielectric* material (e.g. glass). When the electric field is represented using *complex notation*, the index of refraction also becomes a complex number. You should make sure you are comfortable with the material in section 0.2 before proceeding.

2.1 Plane Wave Solutions to the Wave Equation

Consider the wave equation for an electric field waveform propagating in vacuum (1.41):

$$\nabla^2 \mathbf{E} - \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0 \tag{2.1}$$

We are interested in solutions to (2.1) that have the functional form (see P1.9)

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 \cos\left(\mathbf{k} \cdot \mathbf{r} - \omega t + \phi\right) \tag{2.2}$$

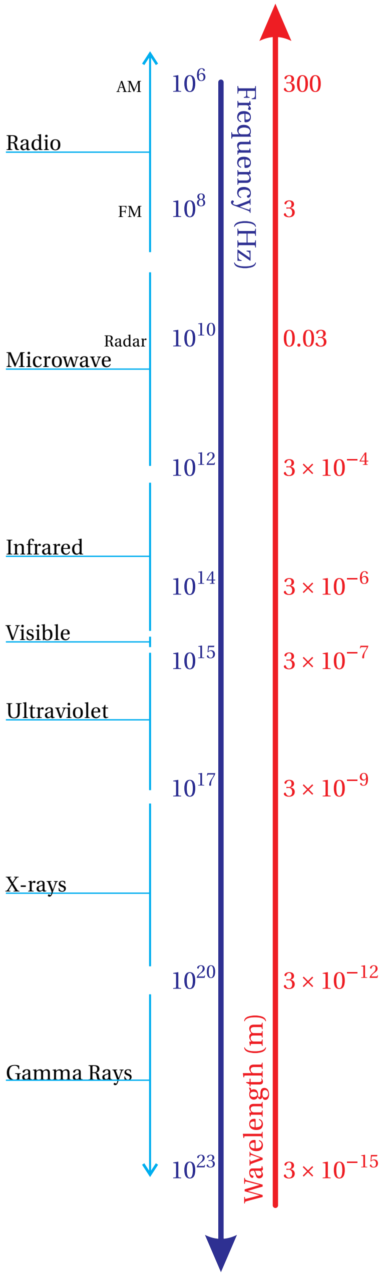


Figure 2.1 The electromagnetic spectrum

Here ϕ represents an arbitrary (constant) phase term. The vector $\mathbf{k}$, called the *wave vector*, may be written as

$$\mathbf{k} \equiv k\hat{\mathbf{u}} = \frac{2\pi}{\lambda_{\text{vac}}}\hat{\mathbf{u}} \qquad \text{(vacuum)} \tag{2.3}$$

where k has units of inverse length, $\hat{\mathbf{u}}$ is a unit vector defining the direction of propagation, and λ_{vac} is the length by which $\mathbf{r}$ must vary (in the direction of $\hat{\mathbf{u}}$) to cause the cosine to go through a complete cycle. This distance is known as the (vacuum) *wavelength*. The *frequency* of oscillation is related to the wavelength via

$$\omega = \frac{2\pi c}{\lambda_{\text{vac}}} \qquad \text{(vacuum)} \tag{2.4}$$

The frequency ω has units of radians per second. Frequency is also often expressed as $\nu \equiv \omega/2\pi$ in units of cycles per second or Hz. Notice that k and ω cannot be chosen independently; the wave equation requires them to be related through the *dispersion relation*

$$k = \frac{\omega}{c} \qquad \text{(vacuum)} \tag{2.5}$$

Typical values for λ_{vac} are given in Fig. 2.1. Sometimes the spatial period of the wave is expressed as $1/\lambda_{\text{vac}}$, in units of cm^{-1}, called the *wave number*.

A magnetic wave accompanies any electric wave, and it obeys a similar wave equation (see P1.6). The magnetic wave corresponding to (2.2) is

$$\mathbf{B}(\mathbf{r}, t) = \mathbf{B}_0 \cos\left(\mathbf{k}\cdot\mathbf{r} - \omega t + \phi\right), \tag{2.6}$$

It is important to note that $\mathbf{B}_0$, $\mathbf{k}$, ω, and ϕ are not independently chosen in (2.6). In order to satisfy Faraday's law (1.3), the arguments of the cosine in (2.2) and (2.6) must be identical. Therefore, in vacuum the electric and magnetic fields travel in phase. In addition, Faraday's law requires (see P1.2)

$$\mathbf{B}_0 = \frac{\mathbf{k} \times \mathbf{E}_0}{\omega} \tag{2.7}$$

The above cross product means that $\mathbf{B}_0$, is perpendicular to both $\mathbf{E}_0$ and $\mathbf{k}$. Meanwhile, Gauss' law $\nabla \cdot \mathbf{E} = 0$ forces $\mathbf{k}$ to be perpendicular to $\mathbf{E}_0$. It follows that the magnitudes of the fields are related through $B_0 = kE_0/\omega$ or $B_0 = E_0/c$, in view of (2.5).

The influence of the magnetic field only becomes important (in comparison to the electric field) for charged particles moving near the speed of light. This typically takes place only for extremely intense lasers ($> 10^{18}$ W/cm^2, see P2.12) where the electric field is sufficiently strong to cause electrons to oscillate with velocities near the speed of light. We will be interested in optics problems that take place at far less intensity where the effects of the magnetic field can typically be safely ignored. Throughout the remainder of this book, we will focus our attention mainly on the electric field with the understanding that we can at any time deduce the (less important) magnetic field from the electric field via Faraday's law.

Figure 2.2 depicts the electric field (2.2) and the associated magnetic field (2.6). The figure is deceptive since the fields don't actually look like transverse waves on a string. The wave is comprised of large planar sheets of uniform field strength (difficult to draw). The name *plane wave* is given since a constant argument in (2.2) at any moment describes a plane, which is perpendicular to **k**. A plane wave fills all space and may be thought of as a series of infinite sheets, each with a different uniform field strength, moving in the **k** direction.

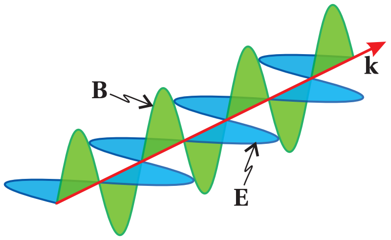


Figure 2.2 Depiction of electric and magnetic fields associated with a plane wave.

2.2 Complex Plane Waves

At this point, let's rewrite our plane wave solution using complex number notation. Although this change in notation will not make the task at hand any easier (and may even appear to complicate things), we introduce the complex notation here in preparation for later sections, where it will save considerable labor. (For a review of complex notation, see section 0.2.)

Using complex notation we rewrite (2.2) as

$$\mathbf{E}(\mathbf{r}, t) = \mathrm{Re}\left\{\tilde{\mathbf{E}}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}\right\} \tag{2.8}$$

where we have hidden the phase term ϕ inside of $\tilde{\mathbf{E}}_0$ as follows:[1]

$$\tilde{\mathbf{E}}_0 \equiv \mathbf{E}_0 e^{i\phi} \tag{2.9}$$

The next step we take is to become intentionally sloppy. Physicists throughout the world have conspired to avoid writing Re{ } in an effort (or lack thereof if you prefer) to make expressions less cluttered. Nevertheless, only the real part of the field is physically relevant even though expressions and calculations contain both real and imaginary terms. This sloppy notation is okay since the real and imaginary parts of complex numbers never intermingle when adding, subtracting, differentiating, or integrating. We can delay taking the real part of the expression until the end of the calculation. Also, when hiding a phase ϕ inside of the field amplitude as in (2.8), we drop the tilde (might as well since we are already being sloppy); we will automatically assume that the field amplitude is complex and contains phase information. Putting this all together, our plane wave solution in complex notation is written simply as

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{2.10}$$

It is possible to construct any electromagnetic disturbance from a linear superposition of such waves, which we will do in chapter 7.

---

[1] We have assumed that each vector component of the field propagates with the same phase. To be more general, one could write $\tilde{\mathbf{E}}_0 \equiv \hat{\mathbf{x}} E_{0x} e^{i\phi_x} + \hat{\mathbf{y}} E_{0y} e^{i\phi_y} + \hat{\mathbf{z}} E_{0z} e^{i\phi_z}$.

Example 2.1

Verify that the complex plane wave (2.10) is a solution to the wave equation (2.1).

Solution: The first term gives

$$\begin{aligned}\nabla^2 \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} &= \mathbf{E}_0\left[\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right]e^{i(k_x x+k_y y+k_z z-\omega t)} \\ &= -\mathbf{E}_0\left(k_x^2+k_y^2+k_z^2\right)e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \\ &= -k^2\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}\end{aligned} \tag{2.11}$$

and the second term gives

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\left(\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}\right) = -\frac{\omega^2}{c^2}\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{2.12}$$

Upon insertion into (2.1) we obtain the vacuum dispersion relation (2.5), which specifies the connection between the wavenumber k and the frequency ω.

2.3 Index of Refraction

Now let's examine how plane waves behave in *dielectric media* (e.g. glass) where electrons are tightly bound to parent atoms or molecules and not free to move about in the material. We assume an *isotropic*,[2] *homogeneous*,[3] and nonconducting medium (i.e. $\mathbf{J}_{\text{free}} = 0$). In this case, we expect **E** and **P** to be parallel to each other so $\nabla\cdot\mathbf{P} = 0$ from (1.33).[4] The general wave equation (1.40) for the electric field reduces in this case to

$$\nabla^2\mathbf{E} - \epsilon_0\mu_0\frac{\partial^2\mathbf{E}}{\partial t^2} = \mu_0\frac{\partial^2\mathbf{P}}{\partial t^2} \tag{2.13}$$

Since we are considering sinusoidal waves, we consider solutions of the form

$$\begin{aligned}\mathbf{E} &= \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \\ \mathbf{P} &= \mathbf{P}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}\end{aligned} \tag{2.14}$$

By writing this, we are making the (reasonable) assumption that if an electric field stimulates a medium at frequency ω, then the polarization in the medium also oscillates at frequency ω. This assumption is typically rather good except for extreme electric fields, which can generate frequency harmonics through nonlinear effects (see P1.12). Recall that by our prior agreement, the complex amplitudes of $\mathbf{E}_0$ and $\mathbf{P}_0$ carry phase information. Thus, while **E** and **P** in (2.14) oscillate at the same frequency, they can be out of phase with respect to each other. This phase discrepancy is most pronounced for materials that absorb energy at the plane wave frequency.

---

[2] Isotropic means the material behaves the same for propagation in any direction. Many crystals are not isotropic as we'll see in Chapter 5.

[3] Homogeneous means the material is everywhere the same throughout space.

[4] This follows for a wave of the form (2.14) if **P** and **k** are perpendicular.

Substitution of the trial solutions (2.14) into (2.13) yields

$$-k^2\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + \epsilon_0\mu_0\omega^2\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} = -\mu_0\omega^2\mathbf{P}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{2.15}$$

To go further, we need to make an explicit connection between $\mathbf{E}_0$ and $\mathbf{P}_0$ (external to Maxwell's equations). In a *linear* medium, the polarization amplitude is proportional to the strength of the applied electric field:

$$\mathbf{P}_0(\omega) = \epsilon_0\chi(\omega)\mathbf{E}_0(\omega) \tag{2.16}$$

This is known as a *constitutive relation.* We have introduced a (to-be-determined) dimensionless proportionality factor $\chi(\omega)$ called the *susceptibility*. We account for the possibility that $\mathbf{E}$ and $\mathbf{P}$ oscillate out of phase by allowing $\chi(\omega)$ to be a complex number. Since $\chi(\omega)$ in general depends on the frequency, we appropriately must think of $\mathbf{P}_0$ and $\mathbf{E}_0$ also as functions of ω.

By inserting (2.16) into (2.15) and canceling the field terms, we obtain the dispersion relation in dielectrics:

$$k^2 = \epsilon_0\mu_0\left[1+\chi(\omega)\right]\omega^2 \qquad \text{or} \qquad k = \frac{\omega}{c}\sqrt{1+\chi(\omega)} \tag{2.17}$$

where we have used $c \equiv 1/\sqrt{\epsilon_0\mu_0}$. The oft-used combination $\epsilon \equiv \epsilon_0(1+\chi)$ is called the permittivity of the material[5]; we will stick with writing out $1+\chi$. In general, $\chi(\omega)$ is a complex number, which leads to a complex *index of refraction*, defined by

$$\mathcal{N}(\omega) \equiv n(\omega) + i\kappa(\omega) = \sqrt{1+\chi(\omega)} \tag{2.18}$$

where n and κ are respectively the real and imaginary parts of the index. (Note that κ is not k.) According to (2.17), the magnitude of the wave vector is also complex according to

$$k = \frac{\mathcal{N}\omega}{c} = \frac{(n+i\kappa)\omega}{c} \tag{2.19}$$

Please keep in mind that the use of a complex index of refraction only makes sense in the context of complex representation for a plane wave.

The complex index $\mathcal{N}$ takes into account absorption as well as the usual oscillatory behavior of the wave. We see this by explicitly placing (2.19) into (2.14):

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0 e^{i(k\hat{\mathbf{u}}\cdot\mathbf{r}-\omega t)} = \mathbf{E}_0 e^{-\frac{\kappa\omega}{c}\hat{\mathbf{u}}\cdot\mathbf{r}} e^{i\left(\frac{n\omega}{c}\hat{\mathbf{u}}\cdot\mathbf{r}-\omega t\right)} \tag{2.20}$$

---

[5] Electrodynamics books often use the electric displacement $\mathbf{D} \equiv \epsilon_0\mathbf{E} + \mathbf{P} = \epsilon\mathbf{E}$. See M. Born and E. Wolf, *Principles of Optics*, 7th ed., p. 3 (Cambridge University Press, 1999). The *permittivity* ϵ encapsulates the constitutive relation that connects $\mathbf{P}$ with $\mathbf{E}$.The index of refraction is given by $\mathcal{N} = \sqrt{\epsilon/\epsilon_0}$.

As before, $\hat{\mathbf{u}}$ is a real unit vector specifying the direction of $\mathbf{k}$. Again, when looking at (2.20), by special agreement in advance, we should just think of the real part, namely

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0 e^{-\frac{\kappa\omega}{c}\hat{\mathbf{u}}\cdot\mathbf{r}} \cos\left(\frac{n\omega}{c}\hat{\mathbf{u}}\cdot\mathbf{r} - \omega t + \phi\right) \tag{2.21}$$

where an overall phase ϕ was formerly held in the complex vector $\tilde{\mathbf{E}}_0$.[6] (The tilde had been suppressed.) Figure 2.3 shows a graph of (2.21). The imaginary part of the complex index (i.e. κ) describes how the wave to decays as it travels. This accounts for absorption. The real part of the complex index (i.e. n) is associated with the oscillations of the wave. By inspection of the cosine argument in (2.21), we see that the speed of the (diminishing) sinusoidal wavefronts is

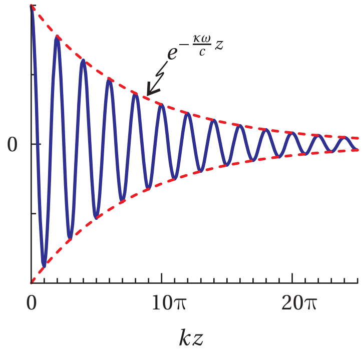


Figure 2.3 Electric field of a decaying plane wave. For convenience in plotting, the direction of propagation is chosen to be in the z direction (i.e. $\hat{\mathbf{u}} = \hat{\mathbf{z}}$).

$$v_{\text{phase}}(\omega) = c/n(\omega) \tag{2.22}$$

so we have $n(\omega) = c/v_{\text{phase}}(\omega)$. Thus, $n(\omega)$ is the ratio of the speed of the light in vacuum to the speed of the wave in the material.

In a dielectric material, the vacuum relations (2.3) and (2.4) are modified to read

$$\text{Re}\{\mathbf{k}\} \equiv \frac{2\pi}{\lambda}\hat{\mathbf{u}}, \tag{2.23}$$

where

$$\lambda \equiv \lambda_{\text{vac}}/n. \tag{2.24}$$

While the frequency ω is the same, whether in a material or in vacuum, the wavelength λ varies with the real part of the complex index.

Example 2.2

When $n = 1.5$, $\kappa = 0.1$, and $\nu = 5 \times 10^{14}$ Hz, find (a) the wavelength inside the material, and (b) the propagation distance over which the amplitude of the wave diminishes by the factor e^{-1} (called the *skin depth*).

Solution:

(a)

$$\lambda = \frac{\lambda_{\text{vac}}}{n} = \frac{2\pi c}{n\omega} = \frac{c}{n\nu} = \frac{\left(3\times 10^8\ \text{m/s}\right)}{1.5\left(5\times 10^{14}\ \text{Hz}\right)} = 400\ \text{nm}$$

(b)

$$e^{-\frac{\kappa\omega}{c}z} = e^{-1} \quad \Rightarrow \quad z = \frac{c}{\kappa\omega} = \frac{c}{2\pi\kappa\nu} = \frac{3\times 10^8\ \text{m/s}}{2\pi(0.1)\left(5\times 10^{14}\ \text{Hz}\right)} = 950\ \text{nm}$$

---

[6] For the sake of simplicity in writing (2.21) we assume linearly polarized light. That is, all vector components of $\tilde{\mathbf{E}}_0$ have the same complex phase ϕ. We will consider other possibilities, such as circularly polarized light, in chapter 6.

Obtaining n and κ from the complex susceptibility χ

From (2.18) we have

$$(n+i\kappa)^2 = n^2 - \kappa^2 + i2n\kappa = 1 + \mathrm{Re}\{\chi\} + i\mathrm{Im}\{\chi\} = 1+\chi \tag{2.25}$$

The real parts and the imaginary parts in the above equation are separately equal:

$$n^2 - \kappa^2 = 1 + \mathrm{Re}\{\chi\} \qquad \text{and} \qquad 2n\kappa = \mathrm{Im}\{\chi\} \tag{2.26}$$

From the latter equation we have

$$\kappa = \mathrm{Im}\{\chi\}/2n \tag{2.27}$$

When this is substituted into the first equation of (2.26) we get a quadratic in n^2

$$n^4 - \left(1+\mathrm{Re}\{\chi\}\right)n^2 - \frac{\left(\mathrm{Im}\{\chi\}\right)^2}{4} = 0 \tag{2.28}$$

The positive[7] real root to this equation is

$$n = \sqrt{\frac{\left(1+\mathrm{Re}\{\chi\}\right) + \sqrt{\left(1+\mathrm{Re}\{\chi\}\right)^2 + \left(\mathrm{Im}\{\chi\}\right)^2}}{2}} \tag{2.29}$$

The imaginary part of the index is then obtained from (2.27).

When absorption is small we can neglect the imaginary part of $\chi(\omega)$, and (2.29) reduces to

$$n(\omega) = \sqrt{1+\chi(\omega)} \qquad \text{(negligible absorption)} \tag{2.30}$$

Hendrik Antoon Lorentz (1853–1928, Dutch) was born in Arnhem, Netherlands, the son a successful nurseryman. Hendrik's mother died when he was four years old. He studied classical languages and then entered the University of Leiden where he was strongly influenced by astronomy professor Frederik Kaiser, whose niece Hendrik married. Hendrik was persuaded to become a physicist and wrote a doctoral dissertation entitled "On the theory of reflection and refraction of light," in which he refined Maxwell's electromagnetic theory. Lorentz correctly hypothesized that the atoms were composed of charged particles, and that their movement was the source of light. He also derived the transformations of space and time, later used in Einstein's theory of relativity. Lorentz won the Nobel prize in 1902 for his contributions to electromagnetic theory. (Wikipedia)

2.4 The Lorentz Model of Dielectrics

To compute the index of refraction in either a dielectric or a conducting material, we require a model that describes the response of electrons in the material to the passing electric field wave. Of course, the model in turn influences how the electric field propagates, which is what influences the material in the first place! The model therefore must be solved together with the propagating field in a self-consistent manner.

Hendrik Lorentz developed a very successful model in the late 1800s, which treats each (active) electron in the medium as a classical particle obeying Newton's second law ($\mathbf{F} = m\mathbf{a}$). In the case of a dielectric medium, electrons are subject to an elastic restoring force that keeps each electron bound to its respective molecule and a damping force that dissipates energy and gives rise to absorption.

---

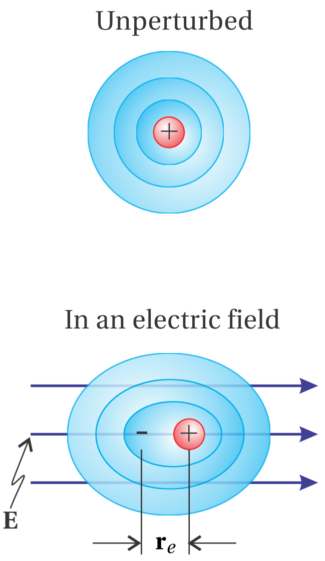
[7] It is possible to have $n < 0$ for so called meta materials, not considered here.

The *Lorentz model* determines the susceptibility $\chi(\omega)$ (the connection between the electric field $\mathbf{E}_0$ and the polarization $\mathbf{P}_0$) and hence the index of refraction. The model assumes that all molecules in the medium are identical, each with one (or a few) active electrons responding to the external field. The atoms are uniformly distributed throughout space with N identical active electrons per volume (units: number per volume). The polarization of the material is then

$$\mathbf{P} = N q_e \mathbf{r}_e \tag{2.31}$$

Recall that polarization has units of dipoles per volume. Each dipole has strength $q_e\mathbf{r}_e$, where $\mathbf{r}_e$ is a microscopic displacement of the electron from equilibrium.

At the time of Lorentz, atoms were thought to be clouds of positive charge wherein point-like electrons sat at rest unless stimulated by an applied electric field. In our modern quantum-mechanical viewpoint, $\mathbf{r}_e$ corresponds to an average displacement of the electronic cloud, which surrounds the nucleus (see Fig. 2.4). The displacement $\mathbf{r}_e$ of the electron charge in an individual atom depends on the *local* strength of the applied electric field $\mathbf{E}$ at the position of the atom. Since the diameter of the electronic cloud is tiny compared to a wavelength of (visible) light, we make the approximation that the electric field is uniform across any individual atom.



Unperturbed

In an electric field

Figure 2.4 A distorted electronic cloud becomes a dipole.

The Lorentz model uses Newton's equation of motion to describe an electron displacement from equilibrium within an atom. In accordance with the classical laws of motion, the electron mass m_e times its acceleration is equal to the sum of the forces on the electron:

$$m_e\ddot{\mathbf{r}}_e = q_e\mathbf{E} - m_e\gamma\dot{\mathbf{r}}_e - k_{\text{Hooke}}\mathbf{r}_e \tag{2.32}$$

The electric field pulls on the electron with force $q_e\mathbf{E}$.[8] A drag force (or friction) $-m_e\gamma\dot{\mathbf{r}}_e$ opposes the electron motion and accounts for absorption of energy. Without this term, it is only possible to describe optical index at frequencies away from where absorption takes place. Finally, $-k_{\text{Hooke}}\mathbf{r}_e$ is a force accounting for the fact that the electron is bound to the nucleus. This restoring force can be thought of as an effective spring that pulls the displaced electron back towards equilibrium with a force proportional to the amount of displacement, so this term is essentially the familiar Hooke's law. With some rearranging, (2.32) can be written as

$$\ddot{\mathbf{r}}_e + \gamma\dot{\mathbf{r}}_e + \omega_0^2\mathbf{r}_e = \frac{q_e}{m_e}\mathbf{E} \tag{2.33}$$

where $\omega_0 \equiv \sqrt{k_{\text{Hooke}}/m_e}$ is the natural oscillation frequency (or resonant frequency) associated with the electron mass and the 'spring constant.'

There is a subtle problem with our analysis, which we will continue to neglect, but which should be mentioned. The field $\mathbf{E}$ in (2.32) is the net field, which is influenced by the presence of all of the dipoles. The actual field that a dipole

---

[8] The electron also experiences a force due to the magnetic field of the light, $\mathbf{F} = q_e\mathbf{v}_e \times \mathbf{B}$, but this force is tiny for typical optical fields.

‘feels’, however, does not include its own field. That is, we should remove from **E** the field produced by each dipole in its own vicinity. This significantly modifies the result if the density of the material is sufficiently high. This effect is described by the *Clausius-Mossotti* formula, which is treated in appendix 2.B.

In accordance with our examination of a single sinusoidal wave, we insert (2.14) into (2.33) and obtain

$$\ddot{\mathbf{r}}_e + \gamma \dot{\mathbf{r}}_e + \omega_0^2 \mathbf{r}_e = \frac{q_e}{m_e} \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{2.34}$$

As a reminder, within a given atom the excursions of $\mathbf{r}_e$ are assumed to be so small that $\mathbf{k}\cdot\mathbf{r}$ remains essentially constant. After all, $\mathbf{k}\cdot\mathbf{r}$ varies typically on a scale of an optical wavelength, which is huge compared to the size of an atom. The inhomogeneous solution to (2.34) is (see P2.1)

$$\mathbf{r}_e = \left(\frac{q_e}{m_e}\right) \frac{\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}}{\omega_0^2 - i\omega\gamma - \omega^2} \tag{2.35}$$

The electron position $\mathbf{r}_e$ oscillates (not surprisingly) with the same frequency ω as the driving electric field. This solution illustrates the convenience of complex notation. The imaginary part in the denominator implies that the electron oscillates with a phase different from the electric field oscillations; the damping term γ (the imaginary part in the denominator) causes the two to be out of phase somewhat. The complex algebra in (2.35) accomplishes quite easily what would otherwise be cumbersome (i.e. working out a trigonometric phase).

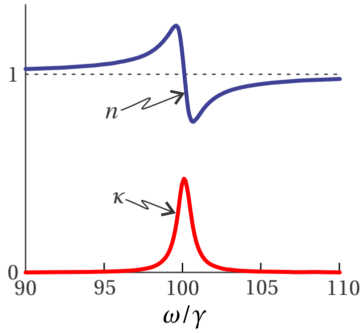


Figure 2.5 Real and imaginary parts of the index for a single Lorentz oscillator dielectric with $\omega_p = 10\gamma$ and $\omega_0 = 100\gamma$.

We are now able to write the polarization in terms of the electric field. By substituting (2.35) into (2.31) and rearranging, we obtain

$$\mathbf{P} = \epsilon_0 \left(\frac{\omega_p^2}{\omega_0^2 - i\omega\gamma - \omega^2}\right) \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{2.36}$$

where the *plasma frequency* ω_p has been introduced:[9]

$$\omega_p \equiv \sqrt{\frac{N q_e^2}{\epsilon_0 m_e}} \tag{2.37}$$

A comparison of (2.36) with (2.16) reveals the (complex) susceptibility:

$$\chi(\omega) = \frac{\omega_p^2}{\omega_0^2 - i\omega\gamma - \omega^2} \tag{2.38}$$

The index of refraction is then found by substituting the susceptibility (2.38) into (2.18). The real and imaginary parts of the index are solved by equating separately the real and imaginary parts of (2.18), namely

$$(n + i\kappa)^2 = 1 + \chi(\omega) = 1 + \frac{\omega_p^2}{\omega_0^2 - i\omega\gamma - \omega^2} \tag{2.39}$$

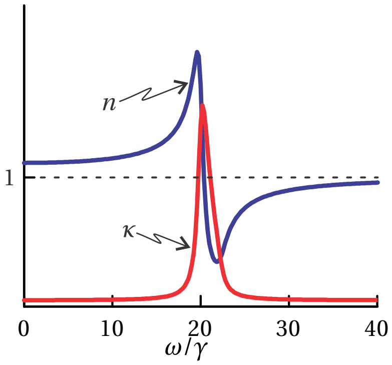


Figure 2.6 Real and imaginary parts of the index for a single Lorentz oscillator dielectric with $\omega_\text{p} = 10\gamma$ and $\omega_0 = 20\gamma$.

Graphs of n and κ are given in Figs. 2.5 and 2.6 for various parameters.

Most materials actually have more than one species of active electron, and different active electrons behave differently. The generalization of (2.39) in this case is

$$(n + i\kappa)^2 = 1 + \chi(\omega) = 1 + \sum_j \frac{f_j \omega_{\text{p}\,j}^2}{\omega_{0\,j}^2 - i\omega\gamma_j - \omega^2} \tag{2.40}$$

where f_j is the aptly named the *oscillator strength* for the j^th species of active electron, inserted into the model without justification to make results better agree with observation. Each species also has its own plasma frequency $\omega_{\text{p}\,j}$, natural frequency $\omega_{0\,j}$, and damping coefficient γ_j. For frequency ranges where $\omega\gamma_j$ and κ can be ignored (i.e. away from resonances $\omega_{0\,j}$), it is common to write Lorentz's refractive-index formula (2.40) in terms of $\lambda_\text{vac} = 2\pi c/\omega$, in which case it is known as the *Sellmeier equation.* (See P2.2.)

Lorentz introduced this model well before the development of quantum mechanics. Even though the model pays no attention to quantum physics, it works surprisingly well for describing frequency-dependent optical indices and absorption of light. As it turns out, the Schrödinger equation applied to two levels in an atom reduces in mathematical form to the Lorentz model in the limit of low-intensity light. Quantum mechanics also explains the oscillator strength, which before the development of quantum mechanics had to be inserted *ad hoc* to make the model agree with experiments. The friction term γ turns out not to be associated with something internal to atoms but rather with collisions between atoms, which on average give rise to the same behavior.

2.5 Index of Refraction of a Conductor

In a conducting medium, the outer electrons of atoms are free to move without being tethered to any particular atom. However, the electrons are still subject to a damping force due to collisions that remove energy and give rise to absorption. Such collisions are associated with resistance in a conductor. As it turns out, we can obtain a simple formula for the refractive index of a conductor from the Lorentz model in section 2.4. We simply remove the restoring force that binds electrons to their atoms. That is, we set $\omega_0 = 0$ in (2.39), which gives

$$(n + i\kappa)^2 = 1 - \frac{\omega_\text{p}^2}{i\omega\gamma + \omega^2} \tag{2.41}$$

This underscores the fact that $\partial\mathbf{P}/\partial t$ is a current very much like $\mathbf{J}_\text{free}$. When we remove the restoring force $k_\text{Hooke} = m_e\omega_0^2$ from the atomic model, the electrons effectively become free, and it is not surprising that they exactly mimic the behavior of a free current $\mathbf{J}_\text{free}$. A graph of n and κ in the conductor model is given

---

[9] In a plasma, charges move freely so that both the Hooke restoring force and the damping term can be neglected (i.e. $\omega_0 \cong 0$, $\gamma \cong 0$). For a plasma, ω_p is the dominant parameter.

in Fig. 2.7. Below, we provide the derivation for (2.41) in the context of $\mathbf{J}_{\text{free}}$ rather than as a limiting case of the dielectric model.[10]

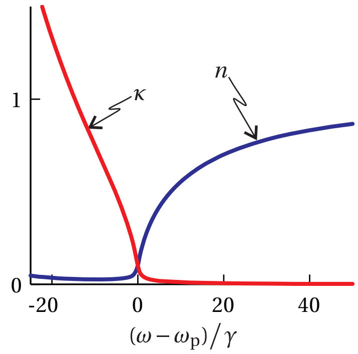


Figure 2.7 Real and imaginary parts of the index for conductor with $\omega_{\text{p}} = 50\gamma$.

Derivation of Refractive Index for a Conductor

We will include the current density $\mathbf{J}_{\text{free}}$ while setting the medium polarization $\mathbf{P}$ to zero. The wave equation (1.40) then becomes

$$\nabla^2\mathbf{E} - \epsilon_0\mu_0\frac{\partial^2}{\partial t^2}\mathbf{E} = \mu_0\frac{\partial}{\partial t}\mathbf{J}_{\text{free}} \tag{2.42}$$

We assume that the current is made up of individual electrons traveling with velocity $\mathbf{v}_e$:

$$\mathbf{J}_{\text{free}} = Nq_{\text{e}}\mathbf{v}_e \tag{2.43}$$

As before, N is the number density of free electrons (in units of number per volume). Recall that current density $\mathbf{J}_{\text{free}}$ has units of charge times velocity per volume (or current per cross sectional area), so (2.43) may be thought of as a definition of current density in a fundamental sense.

Again, the electrons satisfy Newton's equation of motion, similar to (2.32) except without a restoring force:

$$m_e\ddot{\mathbf{r}}_e = q_e\mathbf{E} - m_e\gamma\dot{\mathbf{r}}_e \tag{2.44}$$

For a sinusoidal electric field $\mathbf{E} = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$, the solution to this equation is

$$\mathbf{v}_e \equiv \dot{\mathbf{r}}_e = \left(\frac{q_e}{m_e}\right)\frac{\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}}{\gamma - i\omega} \tag{2.45}$$

where again we assume that the electron oscillation excursions described by $\mathbf{r}_e$ are small compared to the wavelength so that $\mathbf{r}$ can be treated as a constant in (2.44). The current density (2.43) in terms of the electric field is then

$$\mathbf{J}_{\text{free}} = \left(\frac{Nq_e^2}{m_e}\right)\frac{\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}}{\gamma - i\omega} \tag{2.46}$$

We substitute this together with the electric field into the wave equation (2.42) and get

$$-k^2\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + \frac{\omega^2}{c^2}\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} = -i\omega\left(\frac{\mu_0 Nq_e^2}{m_e}\right)\frac{\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}}{\gamma - i\omega} \tag{2.47}$$

This simplifies down to the dispersion relation

$$k^2 = \frac{\omega^2}{c^2}\left(1 - \frac{\omega_{\text{p}}^2}{i\gamma\omega + \omega^2}\right) \tag{2.48}$$

which agrees with (2.41). We have made the substitution $\omega_{\text{p}}^2 = Nq_e^2/\epsilon_0 m_e$ in accordance with (2.37). As usual, $k^2 = \frac{\omega^2(1+\chi)}{c^2} = \frac{\omega^2(n+i\kappa)^2}{c^2}$, so the susceptibility and the index may be extracted from (2.48).

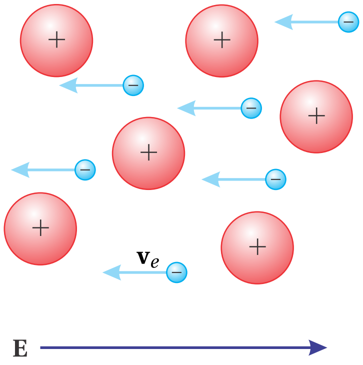


Figure 2.8 The electrons in a conductor can easily move in response to the applied field.

Note that in the low-frequency limit (i.e. $\omega \ll \gamma$), the current density (2.46) reduces to Ohm's law $\mathbf{J} = \sigma \mathbf{E}$, where $\sigma = N q_e^2 / m_e \gamma$ is the DC conductivity. In the high-frequency limit (i.e. $\omega \gg \gamma$), the behavior changes over to that of a free plasma, where collisions, which are responsible for resistance, become less important since the excursions of the electrons during oscillations become very small. This formula captures the general behavior of metals, but actual values of the index vary from this somewhat (see P2.6).

In either the conductor or dielectric model, the damping term removes energy from electron oscillations. The damping term gives rise to an imaginary part of the index, which causes an exponential attenuation of the plane wave as it propagates.

2.6 Poynting's Theorem

Until now, we have described light as the propagation of an electromagnetic disturbance. However, we typically observe light by detecting absorbed energy rather than the field amplitude directly. In this section we examine the connection between propagating electromagnetic fields (such as the plane waves discussed in this chapter) and the energy transported by such fields.

In the late 1800s John Poynting developed (from Maxwell's equations) the theoretical foundation that describes light energy transport. You should appreciate and remember the ideas involved, especially the definition and meaning of the Poynting vector, even if you forget the specifics of its derivation.

John Henry Poynting (1852–1914, English) was the youngest son of a Unitarian minister who operated a school near Manchester, England where John received his childhood education. He later attended Owen's College in Manchester and then went on to Cambridge University where he distinguished himself in mathematics and worked under James Maxwell in the Cavendish Laboratory. Poynting joined the faculty of the University of Birmingham (then called Mason Science College) where he was a professor of physics from 1880 until his death. Besides developing his famous theorem on the conservation of energy in electromagnetic fields, he performed innovative measurements of Newton's gravitational constant and discovered that the Sun's radiation draws in small particles towards it, the Poynting-Robertson effect. Poynting was the principal author of a multi-volume undergraduate physics textbook, which was in wide use until the 1930s. (Wikipedia)

Derivation of Poynting's Theorem

We require just two of Maxwell's Equations: (1.3) and (1.4). We take the dot product of $\mathbf{B}/\mu_0$ with the first equation and the dot product of $\mathbf{E}$ with the second equation. Then by subtracting the second equation from the first we obtain

$$\frac{\mathbf{B}}{\mu_0} \cdot (\nabla \times \mathbf{E}) - \mathbf{E} \cdot \left(\nabla \times \frac{\mathbf{B}}{\mu_0} \right) + \epsilon_0 \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} + \frac{\mathbf{B}}{\mu_0} \cdot \frac{\partial \mathbf{B}}{\partial t} = -\mathbf{E} \cdot \mathbf{J} \tag{2.49}$$

The first two terms can be simplified using the vector identity P0.8. The next two terms are the time derivatives of $\epsilon_0 E^2/2$ and $B^2/2\mu_0$, respectively. The relation (2.49) then becomes

$$\nabla \cdot \left(\mathbf{E} \times \frac{\mathbf{B}}{\mu_0} \right) + \frac{\partial}{\partial t} \left(\frac{\epsilon_0 E^2}{2} + \frac{B^2}{2\mu_0} \right) = -\mathbf{E} \cdot \mathbf{J} \tag{2.50}$$

This is *Poynting's theorem*. Each term in this equation has units of power per volume.

---

[10] G. Burns, *Solid State Physics*, Sect. 9-5 (Orlando: Academic Press, 1985).

It is conventional to write Poynting's theorem as follows:[11]

$$\nabla \cdot \mathbf{S} + \frac{\partial}{\partial t}\left(u_{\text{field}} + u_{\text{medium}}\right) = 0 \tag{2.51}$$

where

$$\mathbf{S} \equiv \mathbf{E} \times \frac{\mathbf{B}}{\mu_0} \tag{2.52}$$

is called the *Poynting vector*, which has units of power per area, called *irradiance*. The expression

$$u_{\text{field}} \equiv \frac{\epsilon_0 E^2}{2} + \frac{B^2}{2\mu_0} \tag{2.53}$$

is the energy per volume stored in the electric and magnetic fields. Derivations of the electric field energy density and the magnetic field energy density are given in Appendices 2.C and 2.D. (See (2.80) and (2.87).) The derivative

$$\frac{\partial u_{\text{medium}}}{\partial t} \equiv \mathbf{E} \cdot \mathbf{J} \tag{2.54}$$

describes the power per volume delivered to the medium from the field. Equation (2.54) is reminiscent of the familiar circuit power law, Power = Voltage × Current. Power is delivered when a charged particle traverses a distance while experiencing a force. This happens when currents flow in the presence of electric fields.

Poynting's theorem is essentially a statement of the conservation of energy, where **S** describes the flow of energy. To appreciate this, consider Poynting's theorem (2.51) integrated over a volume V (enclosed by surface S). If we also apply the divergence theorem (0.11) to the term involving $\nabla \cdot \mathbf{S}$ we obtain

$$\oint_S \mathbf{S} \cdot \hat{\mathbf{n}}\, da = -\frac{\partial}{\partial t} \int_V \left(u_{\text{field}} + u_{\text{medium}}\right) dv \tag{2.55}$$

Notice that the volume integral over energy densities u_{field} and u_{medium} gives the total energy stored in V, whether in the form of electromagnetic field energy density or as energy density that has been given to the medium. The integration of the Poynting vector over the surface gives the net Poynting vector flux directed outward. Equation (2.55) indicates that the outward Poynting vector flux matches the rate that total energy disappears from the interior of V. Conversely, if the Poynting vector is directed inward (negative), then the net inward flux matches the rate that energy increases within V. *The vector* **S** *defines the flow of energy through space.* Its units of *power per area* are just what is needed to describe the brightness of light impinging on a surface.

---

[11] See D. J. Griffiths, *Introduction to Electrodynamics*, 3rd ed., Sect. 8.1.2 (New Jersey: Prentice-Hall, 1999).

Example 2.3

(a) Find the Poynting vector **S** and energy density u_{field} for the plane wave field $\mathbf{E} = \hat{\mathbf{x}}E_0 \cos(kz - \omega t)$ traveling in vacuum. (b) Check that **S** and u_{field} satisfy Poynting's theorem.

Solution: The associated magnetic field is (see P1.2)

$$\mathbf{B} = \frac{\hat{\mathbf{z}}k \times \hat{\mathbf{x}}E_0}{\omega}\cos(kz - \omega t) = \hat{\mathbf{y}}\frac{kE_0}{\omega}\cos(kz - \omega t)$$

(a) The Poynting vector is

$$\begin{aligned}\mathbf{S} = \frac{\mathbf{E} \times \mathbf{B}}{\mu_0} &= \hat{\mathbf{x}}E_0 \cos(kz - \omega t) \times \hat{\mathbf{y}}\frac{kE_0}{\omega\mu_0}\cos(kz - \omega t)\\ &= \hat{\mathbf{z}}c\epsilon_0 E_0^2 \cos^2(kz - \omega t)\end{aligned}$$

where we have used $\omega = kc$ and $\mu_0 = 1/(c^2\epsilon_0)$. The energy density is

$$\begin{aligned}u_{\text{field}} = \frac{\epsilon_0 E^2}{2} + \frac{B^2}{2\mu_0} &= \frac{\epsilon_0 E_0^2}{2}\cos^2(kz - \omega t) + \frac{kE_0^2}{2\mu_0\omega^2}\cos^2(kz - \omega t)\\ &= \epsilon_0 E_0^2 \cos^2(kz - \omega t)\end{aligned}$$

Notice that $S = cu$. The energy density traveling at speed c gives rise to the power per area passing a surface (perpendicular to z).

(b) We have

$$\nabla \cdot \mathbf{S} = c\epsilon_0 E_0^2 \frac{\partial}{\partial z}\cos^2(kz - \omega t) = -2kc\epsilon_0 E_0^2 \cos(kz - \omega t)\sin(kz - \omega t)$$

whereas

$$\frac{\partial u_{\text{field}}}{\partial t} = \epsilon_0 E_0^2 \frac{\partial}{\partial t}\cos^2(kz - \omega t) = 2\omega\epsilon_0 E_0^2 \cos(kz - \omega t)\sin(kz - \omega t)$$

Poynting's theorem (2.50) is satisfied since $\omega = kc$.

It is common to replace the rapidly oscillating function $\cos^2(kz - \omega t)$ with its time average 1/2, but this would have inhibited our ability to take the above derivatives needed in this specific problem.

2.7 Irradiance of a Plane Wave

In this section, we consider the irradiance of a plane wave while propagating in matter. We start with the electric plane-wave field $\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)}$. The magnetic field that accompanies this electric field can be found from Maxwell's equation (1.3), and it turns out to be (compare with problem P1.2)

$$\mathbf{B}(\mathbf{r}, t) = \frac{\mathbf{k} \times \mathbf{E}_0}{\omega} e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} \tag{2.56}$$

When **k** is complex, **B** is out of phase with **E**, and this occurs when absorption takes place. On the other hand, when there is no absorption, then **k** is real, and **B** and **E** carry the same complex phase.

Before computing the Poynting vector (2.52), which involves multiplication, we must remember our unspoken agreement that only the real parts of the fields are relevant. We necessarily remove the imaginary parts before multiplying (see (0.22)). To obtain the real parts of the fields, we add their respective complex conjugates and divide the result by 2 (see (0.30)). The real field associated with the plane-wave electric field is

$$\mathbf{E}(\mathbf{r},t)=\frac{1}{2}\left[\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}+\mathbf{E}_0^* e^{-i(\mathbf{k}^*\cdot\mathbf{r}-\omega t)}\right] \tag{2.57}$$

and the real field associated with (2.56) is

$$\mathbf{B}(\mathbf{r},t)=\frac{1}{2}\left[\frac{\mathbf{k}\times\mathbf{E}_0}{\omega}e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}+\frac{\mathbf{k}^*\times\mathbf{E}_0^*}{\omega}e^{-i(\mathbf{k}^*\cdot\mathbf{r}-\omega t)}\right] \tag{2.58}$$

Now we are ready to calculate the Poynting vector. The algebra is a little messy in general, so we restrict the analysis to the case of an *isotropic medium* for the sake of simplicity.

Calculation of the Poynting Vector for a Plane Wave

Using (2.57) and (2.58) in (2.52) gives

$$\begin{aligned}\mathbf{S}&\equiv\mathbf{E}\times\frac{\mathbf{B}}{\mu_0}\\
&=\frac{1}{2}\left[\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}+\mathbf{E}_0^* e^{-i(\mathbf{k}^*\cdot\mathbf{r}-\omega t)}\right]\times\frac{1}{2\mu_0}\left[\frac{\mathbf{k}\times\mathbf{E}_0}{\omega}e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}+\frac{\mathbf{k}^*\times\mathbf{E}_0^*}{\omega}e^{-i(\mathbf{k}^*\cdot\mathbf{r}-\omega t)}\right]\\
&=\frac{1}{4\mu_0}\left[\begin{array}{l}\frac{\mathbf{E}_0\times(\mathbf{k}\times\mathbf{E}_0)}{\omega}e^{2i(\mathbf{k}\cdot\mathbf{r}-\omega t)}+\frac{\mathbf{E}_0^*\times(\mathbf{k}\times\mathbf{E}_0)}{\omega}e^{i(\mathbf{k}-\mathbf{k}^*)\cdot\mathbf{r}}\\
\quad+\frac{\mathbf{E}_0\times\left(\mathbf{k}^*\times\mathbf{E}_0^*\right)}{\omega}e^{i(\mathbf{k}-\mathbf{k}^*)\cdot\mathbf{r}}+\frac{\mathbf{E}_0^*\times\left(\mathbf{k}^*\times\mathbf{E}_0^*\right)}{\omega}e^{-2i(\mathbf{k}^*\cdot\mathbf{r}-\omega t)}\end{array}\right]\end{aligned} \tag{2.59}$$

Very often, we are interested in the time-average of the Poynting vector, denoted by $\langle\mathbf{S}\rangle_t$, since there are no electronics that can keep up with the rapid oscillation of visible light (i.e. $> 10^{14}$ Hz). The first and last terms in (2.59) rapidly oscillate and vanish under time averaging.

Additionally, we can use the BAC-CAB rule P0.3 to write $\mathbf{E}_0^*\times(\mathbf{k}\times\mathbf{E}_0)=\mathbf{k}\left(\mathbf{E}_0^*\cdot\mathbf{E}_0\right)$ and similarly $\mathbf{E}_0\times\left(\mathbf{k}^*\times\mathbf{E}_0^*\right)=\mathbf{k}^*\left(\mathbf{E}_0\cdot\mathbf{E}_0^*\right)$, where we have employed $\mathbf{k}\cdot\mathbf{E}_0=0$, which follows from $\nabla\cdot\mathbf{E}=0$ in an isotropic medium (i.e. not a crystal). The time-averaged Poynting vector then reduces to

$$\langle\mathbf{S}\rangle_t=\frac{\mathbf{k}+\mathbf{k}^*}{4\mu_0\omega}\left(\mathbf{E}_0\cdot\mathbf{E}_0^*\right)e^{i(\mathbf{k}-\mathbf{k}^*)\cdot\mathbf{r}} \qquad \text{(isotropic medium)} \tag{2.60}$$

We can further simplify this expression using $\mathbf{k}=\hat{\mathbf{u}}\,(n+i\kappa)\,\omega/c$ (see (2.19)). We can also use (1.42) to rewrite $1/\mu_0 c$ as $\epsilon_0 c$, in which case (2.60) becomes

$$\langle \mathbf{S} \rangle_t = \hat{\mathbf{u}} \frac{n\epsilon_0 c}{2} \left(\mathbf{E}_0 \cdot \mathbf{E}_0^* \right) e^{-2\frac{\kappa\omega}{c}\hat{\mathbf{u}}\cdot\mathbf{r}} \qquad \text{(isotropic medium)} \quad (2.61)$$

This expression shows that (in an isotropic medium) the flow of energy is in the direction of $\hat{\mathbf{u}}$ (or $\mathbf{k}$). This agrees with our intuition that energy flows in the direction that the wave propagates.

The magnitude of expression (2.61) is the irradiance. However, we often refer to it as the intensity of a field I, which amounts to the same thing, but without regard for the flow of energy. The definition of intensity is thus less specific, and it can be applied, for example, to standing waves where the net Poynting flux of counter-propagating plane waves is technically zero since the two plane waves have equal amounts of energy, but propagate in opposite directions. Nevertheless, atoms in standing waves 'feel' the oscillating field, and we ascribe an intensity to it.

In general, the intensity is written as

$$I = \frac{n\epsilon_0 c}{2} \mathbf{E}_0 \cdot \mathbf{E}_0^* = \frac{n\epsilon_0 c}{2} \left(|E_{0x}|^2 + \left|E_{0y}\right|^2 + |E_{0z}|^2 \right) \qquad (2.62)$$

where in this case we have ignored absorption (i.e. $\kappa \approx 0$). Alternatively, we could consider $|E_{0x}|^2$, $\left|E_{0y}\right|^2$, and $|E_{0z}|^2$ to include the factor $\exp(-2(\kappa\omega/c)\hat{\mathbf{u}}\cdot\mathbf{r})$ so that they correspond to the *local* electric field. Equation (2.62) agrees with S in Example 2.3 where $n = 1$ and $\mathbf{E}_0 = \hat{\mathbf{x}} E_0$ is real; the cosine squared averages to $1/2$.

Radiant Power (of a source): Electromagnetic energy per unit time. Units: $\text{W} = \text{J/s}$

Radiant Solid-Angle Intensity (of a source): Radiant power per steradian emitted from a point-like source (4π steradians in a sphere). Units: W/Sr

Radiance or Brightness (of a source): Radiant solid-angle intensity per unit projected area of an extended source. The *projected* area foreshortens by $\cos\theta$, where θ is the observation angle relative to the surface normal. Units: $\text{W}/(\text{Sr}\cdot\text{cm}^2)$

Radiant Emittance or Exitance (from a source): Radiant Power emitted per unit surface area of an extended source (the Poynting flux leaving). Units: W/cm^2

Irradiance (to a receiver), often called intensity: Electromagnetic power delivered per area to a receiver: Poynting flux arriving. Units: W/cm^2

Table 2.1 Radiometric quantities and units.

Appendix 2.A Radiometry, Photometry, and Color

Radiometry

The field of study that quantifies the energy in electromagnetic radiation (including visible light) is referred to as *radiometry*. Table 2.1 lists several concepts important in radiometry. The radiance at a detector and the exitance from a source are both direct measurements of the average Poynting flux, and the other quantities in the table are directly related to the Poynting flux through geometric factors. One of the challenges in radiometry is that light sensors always have a wavelength-dependent sensitivity to light, whereas the quantities in Table 2.1 treat light of all wavelengths on equal footing. Disentangling the detector response from the desired signal in a radiometric measurement takes considerable care.

Photometry

Photometry refers to the characterization of light energy in the context of the response of the human eye. In contrast to radiometry, photometry takes great care to mimic the wavelength-dependent effects of the eye-brain detection system so that photometric quantities are an accurate reflection of our everyday experience with light. The concepts used in photometry are similar to those in radiometry, except that the radiometric quantities are multiplied by the spectral response of our eye-brain system.

Our eyes contain two types of photoreceptors—rods and cones. The rods are very sensitive and provide virtually all of our vision in dim light conditions. Under these conditions we experience *scotopic vision*, with a response curve that peaks at $\lambda_{\text{vac}} = 507$ nm and is insensitive to wavelengths longer than 640 nm[12] (see Fig. 2.9). As the light gets brighter the less-sensitive cones take over, and we experience *photopic vision*, with a response curve that peaks at $\lambda_{\text{vac}} = 555$ nm and drops to near zero for wavelengths longer than $\lambda_{\text{vac}} = 700$ nm or shorter than $\lambda_{\text{vac}} = 400$ nm (see Fig. 2.9). Photometric quantities are usually measured using the bright-light (photopic) response curve since that is what we typically experience in normally lit spaces.

The basic unit of luminous power is called the *lumen*, defined to be (1/683) W of light with wavelength $\lambda_{\text{vac}} = 555$ nm, the peak of the eye's bright-light response. At other wavelengths, additional radiant power is required to achieve the same number of lumens, as plotted in Fig. 2.9. Photometric units were first defined by comparing sources to burning candles with prescribed dimensions made from whale tallow. Today, the procedure for measuring luminous power is essentially to measure the radiometric power spectrum $I(\lambda)$, and then calculate

$$\text{Lumens} = \int R(\lambda) I(\lambda) d\lambda \tag{2.63}$$

where $R(\lambda)$ is the photopic photometric response function plotted in Fig. 2.9.

Photometric units are often used to characterize room lighting as well as photographic, projection, and display equipment. For example, both a 60 W incandescent bulb and a 13 W compact fluorescent bulb emit about 800 lumens of light, while their radiometric output is much closer to their power rating. The difference in photometric output versus radiometric output reflects the fact that most of the energy radiated from an incandescent bulb is emitted in the infrared, where our eyes are not sensitive. Table 2.2 gives the names of the various photometric quantities, which parallel the entries for radiometric quantities in Table 2.1.

Color

In addition to brightness, our eye-brain measures some basic information about the spectral content of light. We interpret this spectral information as the color of the light. Color information arises from the cone receptors in the eye, which come in three varieties, each sensitive to light in a different wavelength band.

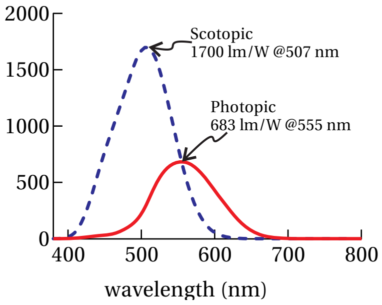


Figure 2.9 The response of a "standard" human eye under relatively bright conditions (photopic) and in dim conditions (scotopic).

Luminous Power (of a source): Visible light energy emitted per time from a source. Units: lumens (lm) lm=(1/683) W @ 555 nm

Luminous Solid-Angle Intensity (of a source) Luminous power per steradian emitted from a point-like source. Units: candelas (cd), cd = lm/Sr.

Luminance (of a source): Luminous solid-angle intensity per projected area of an extended source. (The projected area foreshortens by $\cos\theta$, where θ is the observation angle relative to the surface normal.) Units: cd/cm^2 = stilb, cd/m^2 = nit, nit = 3183 lambert = 3.4 footlambert

Luminous Emittance or Exitance (from a source): Luminous Power emitted per unit surface area of an extended source. Units: lm/cm^2

Illuminance (to a receiver): Incident luminous power delivered per area to a receiver. Units: lux; lm/m^2 = lux, lm/cm^2 = phot, lm/ft^2 = footcandle

Table 2.2 Photometric quantities and units.

---

[12] Since rods do not detect the longer red wavelengths, it is possible to have artificial red illumination without ruining your dark-adapted vision. For example, an airplane can have red illumination on the instrument panel without interfering with a pilot's ability to achieve full dark-adapted vision to see things outside the cockpit.

Figure 2.10 plots the normalized sensitivity curves[13] for short (S), medium (M), and long (L) wavelength cones. When the three types of cones are stimulated equally the light appears white, and when they are stimulated differently the light appears colored.

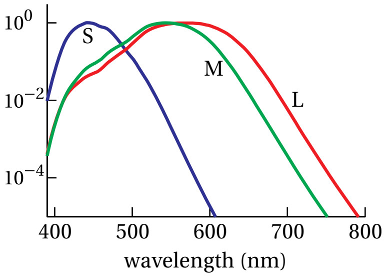


Figure 2.10 Normalized cone sensitivity functions

Light with different spectral distributions can produce the exact same color sensation, so our perception of color only gives very general information about the spectral content of light. For example, light coming from a computer display has a different spectral composition than the light incident on the camera that recorded the image, but both can produce the same color sensation. This ambiguity can lead to a potentially dangerous situation in the lab. For example, lasers from 670 nm to 800 nm all appear the same color. (They all stimulate the L and M cones in essentially the same ratio.) However, your eye's response falls off quickly in the near-infrared, so a dangerous 800 nm high-intensity beam can appear about the same brightness as an innocuous 670 nm laser pointer.

Because we have three types of cones, our perception of color can be well-represented using a three-dimensional vector space referred to as a *color space*.[14] A color space can be defined in terms of three "basis" light sources referred to as *primaries*. Different colors (i.e. the "vectors" in the color space) are created by mixing the primary light in different ratios. If we had three primaries that separately stimulated each type of rod (S, M, and L), we could recreate any color sensation exactly by mixing those primaries. However, by inspecting Fig. 2.10 you can see that this ideal set of primaries cannot be found because of the overlap between the S, M, and L curves. Any light that will stimulate one type of cone will also stimulate another. This overlap makes it impossible to *display* every possible color with three primaries. However, it *is* possible to quantify all colors with three primaries, even if the primaries can't display the colors—we'll see how shortly. The range of colors that can be displayed with a given set of primaries is referred to as the *gamut* of that color space. As your experience with computers suggests, we are able to engineer devices with a very broad gamut, but there are always colors visible in nature that cannot be recreated by a three-primary device.

The CIE1931 RGB[15] color space is a very commonly encountered color space based on a series of experiments performed by W. David Wright and John Guild in the late 1920s. In these experiments, test subjects were asked to match the color of a monochromatic test light source by mixing monochromatic primaries at 700 nm (R), 546.1 nm (G), and 435.8 nm (B). The relative amount of R, G, and B light required to match the color at each test wavelength was recorded as the *color matching functions* $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, and $\bar{b}(\lambda)$, shown in Fig. 2.11. Note that the color matching functions sometimes go negative. This is most noticeable for $\bar{r}(\lambda)$, but all three have negative values. These negative values indicate that the test color was outside the gamut of the primaries (i.e. the color of the test source could not be matched by adding primaries). In these cases, the observers matched the test light as closely as possible by mixing primaries, and then they added some of the primary light to the *test* light until the colors matched. The amount of primary light that had to be added to the test light was recorded as a negative number. In this way they were able to quantify the color, even though it couldn't be displayed using their primaries.

---

[13] A. Stockman, L. Sharpe, and C. Fach, "The spectral sensitivity of the human short-wavelength cones," *Vision Research*, **39**, 2901-2927 (1999); A. Stockman, and L. Sharpe, "Spectral sensitivities of the middle- and long-wavelength sensitive cones derived from measurements in observers of known genotype," *Vision Research*, **40**, 1711-1737 (2000).

[14] The methods we use to represent color are very much tied to human physiology. Other species have photoreceptors that sense different wavelength ranges or do not sense color at all. For instance, Papilio butterflies have six types of cone-like photoreceptors and certain types of shrimp have twelve. Reptiles have four-color vision for visible light, and pit vipers have an additional set of "eyes" that look like pits on the front of their face. These pits are essentially pinhole cameras sensitive to infrared light, and give these reptiles crude night-vision capabilities. On the other hand, some insects can perceive markings on flowers that are only visible in the ultraviolet. Each of these species would find the color spaces we use to record and display colors to be very inaccurate.

[15] CIE is an abbreviation for the French "Commission Internationale de l'Éclairage," an international commission that defines lighting and color standards. This standard was adopted in 1931, and hence the name. Note that CIE1931 is not the RGB space most commonly encountered on a computer to define colors on webpages and in photos—that space is referred to as sRGB and uses a

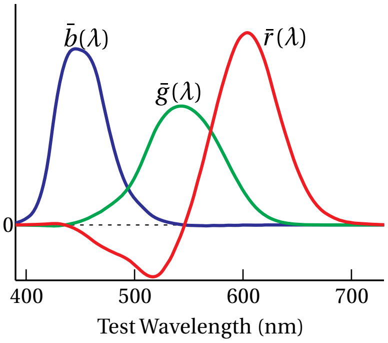


Figure 2.11 The CIE 1931 RGB color-matching functions.

To calculate the color components of an arbitrary light source with a radiometric spectrum $I(\lambda)$, we integrating the spectrum against the color matching functions:

$$R = \int I(\lambda)\bar{r}\,d\lambda \qquad G = \int I(\lambda)\bar{g}\,d\lambda \qquad B = \int I(\lambda)\bar{b}\,d\lambda \tag{2.64}$$

The triplet of numbers (R, G, B) then uniquely define the color of the light source. If R, G, or B turn out to be negative for a given $I(\lambda)$, then that color of light falls outside the gamut of these particular primaries. However, the negative coordinates still provide a valid abstract representation of that color.

The RGB color space is an *additive* color model, where light emitting primaries are added together to produce color and the absence of light gives black. *Subtractive* color models produce color starting with a white reflective substrate (i.e. something that reflects all frequencies of visible light equally like a piece of paper or canvas) and then placing absorbing pigments over the substrate to remove portions of the reflected spectrum.

Some schemes for displaying colors employ more than three basis vectors. For example, color printers typically use the subtractive cyan, magenta, yellow, and black (CMYK) color space, and some display manufacturers add a fourth additive primary, such as yellow, to the typical set of red, green, and blue primaries. The extra basis vector increases the range of colors that can be *displayed* by these systems (i.e. it increases the gamut). However, the fourth basis vector makes the color space overdetermined and only helps in displaying colors—we can abstractly represent all colors using just three coordinates in an appropriately chosen basis.

Example 2.4

---

different set of primaries.

The CIE1931 XYZ color space is derived from the CIE1931 RGB space by the transformation

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \frac{1}{0.17697} \begin{bmatrix} 0.49 & 0.31 & 0.20 \\ 0.17697 & 0.81240 & 0.01063 \\ 0.00 & 0.01 & 0.99 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.65}$$

where X, Y, and Z are the color coordinates in the new basis. The matrix elements in (2.65) were carefully chosen to give this color space some desirable properties: the new coordinates (X, Y, or Z) are always positive; the Y coordinate gives the photometric brightness of the light while the X and Z coordinates describe the color part (i.e. the *chromaticity*) of the light; and the coordinate $(1/3,1/3,1/3)$ gives the color white.

The XYZ coordinates do not represent new primaries, but rather linear combinations of the original primaries. Find the representation in the CIE1931 RGB basis for each of the basis vectors in the XYZ space.

Solution: We first invert the transformation matrix to find

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 0.4185 & -0.1587 & -0.08283 \\ -0.09117 & 0.2524 & 0.01571 \\ 0.0009209 & -0.002550 & 0.1786 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

For the X basis vector, the RGB components are found as

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 0.4185 & -0.1587 & -0.08283 \\ -0.09117 & 0.2524 & 0.01571 \\ 0.0009209 & -0.002550 & 0.1786 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Thus, the X basis vector RGB components are $(0.4185, -0.09117, +0.0009209)$. Similar calculations for the Y and Z basis vectors give $(-0.1587, 0.2524, -0.002550)$ and $(-0.08283, 0.01571, 0.1786)$, respectively. Because the XYZ basis vectors contain negative amounts of the physical RGB primaries, the XYZ basis is not physically realizable. However, it is extensively used because it can abstractly represent all colors using a triplet of positive numbers.

Appendix 2.B Clausius-Mossotti Relation

Equation (2.35) has the form $\mathbf{r}_e = \alpha\mathbf{E}/q_e$, where α is called the atomic (or molecular) *polarizability*. We take absorption to be negligible so that α is real. **E** is the macroscopic field in the medium, which includes a contribution from all of the dipoles. To avoid double-counting the dipole's own field, we should replace **E** with

$$\mathbf{E}_{\text{actual}} \equiv \mathbf{E} - \mathbf{E}_{\text{dipole}} \tag{2.66}$$

and write

$$q_e\mathbf{r}_e = \alpha\mathbf{E}_{\text{actual}} \tag{2.67}$$

That is, we ought not to allow the dipole's own field to act on itself as we previously (inadvertently) did. Here $\mathbf{E}_{\text{dipole}}$ is the average field that a dipole contributes to its quota of space in the material.

Since N is the number of dipoles per volume, each dipole occupies a volume $1/N$. As will be shown below, the average field due to a dipole[16] centered in such a volume (symmetrically chosen) is

$$\mathbf{E}_{\text{dipole}} = -\frac{N q_e \mathbf{r}_\text{e}}{3\epsilon_0} \tag{2.68}$$

Substitution of (2.68) and (2.67) into (2.66) yields

$$\mathbf{E}_{\text{actual}} = \mathbf{E} + \frac{N\alpha \mathbf{E}_{\text{actual}}}{3\epsilon_0} \quad \Rightarrow \quad \mathbf{E}_{\text{actual}} = \frac{\mathbf{E}}{1 - \frac{N\alpha}{3\epsilon_0}} \tag{2.69}$$

Then (2.67) becomes

$$q_e \mathbf{r}_e = \frac{\alpha \mathbf{E}}{1 - \frac{N\alpha}{3\epsilon_0}} \tag{2.70}$$

According to (2.16), the susceptibility is defined via $\mathbf{P} = \epsilon_0 \chi \mathbf{E}$, where $\mathbf{E}$ is the macroscopic field. The polarization is always based on the combined behavior of all of the dipoles $\mathbf{P} = N q_e \mathbf{r}_\text{e}$ (see (2.31)). Equating these two expressions for $\mathbf{P}$ and inserting (2.70), we find that the susceptibility is given by

$$\chi(\omega) = \frac{\frac{N\alpha(\omega)}{\epsilon_0}}{1 - \frac{N\alpha(\omega)}{3\epsilon_0}} \tag{2.71}$$

This is known as the Clausius-Mossotti relation. In Section 2.4, we only included the numerator of (2.71). The extra term in the denominator becomes important when N is sufficiently large, which is the case for liquid or solid densities.

Since we neglect absorption, from (2.25) we have $\chi = n^2 - 1$, and we may write

$$n^2 - 1 = \frac{N\alpha/\epsilon_0}{1 - N\alpha/3\epsilon_0} \tag{2.72}$$

In this case, we may invert the relation to write $N\alpha/\epsilon_0$ in terms of the index:[17]

$$\frac{N\alpha}{\epsilon_0} = 3\frac{n^2 - 1}{n^2 + 2} \tag{2.73}$$

---

[16] In principle, the detailed fields of nearby dipoles should also be considered rather than representing their influence with the macroscopic field. However, if they are symmetrically distributed the result is the same. See J. D. Jackson, *Classical Electrodynamics*, 3rd ed., Sect. 4.5 (New York: John Wiley, 1999).

[17] This form of Clausius-Mossotti relation, in terms of the refractive index, was renamed the Lorentz-Lorenz formula, but probably undeservedly so, since it is essentially the same formula.

Example 2.5

Xenon vapor at STP (density $4.46\times10^{-5}\mathrm{mol/cm^3}$) has index $n = 1.000702$ measured at wavelength 589nm. Use (a) the Clausius-Mossotti relation (2.71) and (b) the uncorrected formula (i.e. numerator only) to predict the index for liquid xenon with density $2.00\times10^{-2}\mathrm{mol/cm^3}$. Compare with the measured value of $n = 1.332$.[18]

Solution: At the low density, we may safely neglect the correction in the denominator of (2.72) and simply write $N_{\text{atm}}\alpha/\epsilon_0 = 1.000702^2 - 1 = 1.404\times10^{-3}$. The liquid density N_{liquid} is $2.00\times10^{-2}/4.46\times10^{-5} = 449$ times greater. Therefore, $N_{\text{liquid}}\alpha/\epsilon_0 = 449\times1.404\times10^{-3} = 0.630$. (a) According to Clausius-Mossotti (2.72), the index is

$$n = \sqrt{1 + \frac{0.630}{1 - 0.630/3}} = 1.341$$

(b) On the other hand, without the correction in the denominator, we get

$$n = \sqrt{1 + 0.630} = 1.277$$

The Clausius-Mossotti formula gets much closer to the measured value.

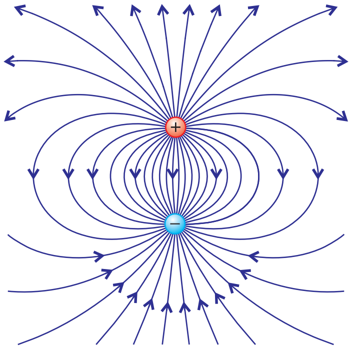


Figure 2.12 The field lines surrounding a dipole.

Average Field Produced by a Dipole

Consider a dipole comprised of point charges $\pm q_e$ separated by spacing $\mathbf{r}_e = \hat{\mathbf{z}}d$. If the dipole is centered on the origin, then by Coulomb's law the field surrounding the point charges is

$$\mathbf{E} = \frac{q_e}{4\pi\epsilon_0}\frac{\mathbf{r}-\hat{\mathbf{z}}d/2}{|\mathbf{r}-\hat{\mathbf{z}}d/2|^3} - \frac{q_e}{4\pi\epsilon_0}\frac{\mathbf{r}+\hat{\mathbf{z}}d/2}{|\mathbf{r}+\hat{\mathbf{z}}d/2|^3}$$

We wish to compute the average field within a cubic volume $V = L^3$ that symmetrically encompasses the dipole.[19] We take the volume dimension L to be large compared to the dipole dimension d. Integrating the field over this volume yields

$$\begin{aligned}\int \mathbf{E}\,dv &= \frac{q_e}{4\pi\epsilon_0}\int_{-L/2}^{L/2}dx\int_{-L/2}^{L/2}dy\int_{-L/2}^{L/2}dz\left[\frac{x\hat{\mathbf{x}}+y\hat{\mathbf{y}}+(z-d/2)\hat{\mathbf{z}}}{\left[x^2+y^2+(z-d/2)^2\right]^{3/2}} - \frac{x\hat{\mathbf{x}}+y\hat{\mathbf{y}}+(z+d/2)\hat{\mathbf{z}}}{\left[x^2+y^2+(z+d/2)^2\right]^{3/2}}\right]\\ &= -\hat{\mathbf{z}}\frac{q_e}{2\pi\epsilon_0}\int_{-L/2}^{L/2}dx\int_{-L/2}^{L/2}dy\left[\frac{1}{\sqrt{x^2+y^2+(L-d)^2/4}} - \frac{1}{\sqrt{x^2+y^2+(L+d)^2/4}}\right]\end{aligned}$$

---

[18] D. H. Garside, H. V. Molgaard, and B. L. Smith, "Refractive Index and Lorentz-Lorenz function of Xenon Liquid and Vapour," J. Phys. B: At. Mol. Phys. **1**, 449-457 (1968).

[19] Authors often obtain the same result using a spherical volume with the (usually unmentioned) conceptual awkwardness that spheres cannot be closely packed to form a macroscopic medium without introducing voids.

The terms multiplying $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ vanish since they involve odd functions integrated over even limits on either x or y, respectively. On the remaining term, the integration on z has been executed. Before integrating the remaining expression over x and y, we make the following approximation based on $L >> d$:

$$\begin{aligned}\frac{1}{\sqrt{x^2+y^2+(L\pm d)^2/4}} &\cong \frac{1}{\sqrt{x^2+y^2+L^2/4}}\frac{1}{\sqrt{1\pm\frac{Ld/2}{x^2+y^2+L^2/4}}} \\ &\cong \frac{1}{\sqrt{x^2+y^2+L^2/4}}\left[1\mp\frac{Ld/4}{x^2+y^2+L^2/4}\right]\end{aligned}$$

which will make integration considerably easier.[20] Then integration over the y dimension brings us to[21]

$$\int \mathbf{E}dv = -\hat{\mathbf{z}}\frac{q_e d}{4\pi\epsilon_0}\int_{\text{-L/2}}^{\text{L/2}} dx\int_{\text{-L/2}}^{\text{L/2}}\frac{Ldy}{\left[x^2+y^2+L^2/4\right]^{3/2}} = -\hat{\mathbf{z}}\frac{q_e d}{4\pi\epsilon_0}\int_{\text{-L/2}}^{\text{L/2}}\frac{L^2dx}{\left(x^2+L^2/4\right)\sqrt{x^2+L^2/2}}$$

The final integral is the same as twice the integral from 0 to $L/2$. Then, with $x > 0$, we can employ the variable change $s = x^2+L^2/4 \Rightarrow 2dx = ds/\sqrt{s-L^2/4}$ and obtain

$$\int \mathbf{E}dv = -\hat{\mathbf{z}}\frac{q_e d}{4\pi\epsilon_0}\int_{L^2/4}^{L^2/2}\frac{L^2ds}{s\sqrt{s^2-L^4/16}} = -\hat{\mathbf{z}}\frac{q_e d}{4\pi\epsilon_0}\frac{4\pi}{3}$$

Reinstalling $\mathbf{r}_\mathrm{e} = \hat{\mathbf{z}}d$ and dividing by the volume $1/N$, allotted to individual dipoles, brings us to the anticipated result (2.68).

Appendix 2.C Energy Density of Electric Fields

In this appendix we show that the term $\epsilon_0 E^2/2$ in (2.53) corresponds to the energy density of an electric field.[22] The electric potential $\phi(\mathbf{r})$ (in units of energy per

---

[20] One might be tempted to begin this calculation with the well-known dipole field

$$\mathbf{E} = \frac{q_e}{4\pi\epsilon_0 r^3}\left[\frac{\mathbf{r}-\hat{\mathbf{z}}d/2}{\left[1-\hat{\mathbf{z}}\cdot\hat{\mathbf{r}}\frac{d}{r}+\frac{d^2}{4r^2}\right]^{3/2}} - \frac{\mathbf{r}+\hat{\mathbf{z}}d/2}{\left[1+\hat{\mathbf{z}}\cdot\hat{\mathbf{r}}\frac{d}{r}+\frac{d^2}{4r^2}\right]^{3/2}}\right] \cong \frac{q_e d}{4\pi\epsilon_0 r^3}\left[3\hat{\mathbf{r}}\left(\hat{\mathbf{z}}\cdot\hat{\mathbf{r}}\right)-\hat{\mathbf{z}}\right]$$

which relies on the approximation

$$\left[1\pm\hat{\mathbf{z}}\cdot\hat{\mathbf{r}}d/r+d^2/4r^2\right]^{-3/2} \cong \left[1\pm\hat{\mathbf{z}}\cdot\hat{\mathbf{r}}d/r\right]^{-3/2} \cong 1\mp\frac{3d\hat{\mathbf{z}}\cdot\hat{\mathbf{r}}}{2r}$$

This dipole-field expression, while useful for describing the field surrounding the dipole, contains no information about the fields internal to the dipole. Note that we integrate z through the origin, which would violate the above assumption $r \gg d$. Alternatively, the influence of the internal fields on our integral could be accomplished using a delta function as is done in J. D. Jackson, *Classical Electrodynamics*, 3rd ed., p. 149 (New York: John Wiley, 1999).

[21] Two useful integral formulas are (0.61) and (0.61).

[22] J. R. Reitz, F. J. Milford, and R. W. Christy, *Foundations of Electromagnetic Theory* 3rd ed., Sect. 6-3 (Reading, Massachusetts: Addison-Wesley, 1979).

charge, or volts) describes the potential energy that a charge would experience if placed at any given point in the field. The electric field and the potential are connected through

$$\mathbf{E}(\mathbf{r}) = -\nabla\phi(\mathbf{r}) \tag{2.74}$$

The energy U necessary to assemble a distribution of charges (owing to attraction or repulsion) can be written in terms of a summation over all of the charges (or charge density $\rho(\mathbf{r})$) located within the potential:

$$U = \frac{1}{2}\int\limits_V \phi(\mathbf{r})\,\rho(\mathbf{r})\,dv \tag{2.75}$$

We consider the potential to arise from the charges themselves. The factor 1/2 is necessary to avoid double counting. To appreciate this factor consider just two point charges: We only need to count the energy due to one charge in the presence of the other's potential to obtain the energy required to bring the charges together.

A substitution of (1.1) for $\rho(\mathbf{r})$ into (2.75) gives

$$U = \frac{\epsilon_0}{2}\int\limits_V \phi(\mathbf{r})\,\nabla\cdot\mathbf{E}(\mathbf{r})\,dv \tag{2.76}$$

Next, we use the vector identity in P0.9 and get

$$U = \frac{\epsilon_0}{2}\int\limits_V \nabla\cdot\left[\phi(\mathbf{r})\,\mathbf{E}(\mathbf{r})\right]dv - \frac{\epsilon_0}{2}\int\limits_V \mathbf{E}(\mathbf{r})\cdot\nabla\phi(\mathbf{r})\,dv \tag{2.77}$$

An application of the divergence theorem (0.11) on the first integral and a substitution of (2.74) into the second integral yields

$$U = \frac{\epsilon_0}{2}\oint\limits_S \phi(\mathbf{r})\,\mathbf{E}(\mathbf{r})\cdot\hat{\mathbf{n}}\,da + \frac{\epsilon_0}{2}\int\limits_V \mathbf{E}(\mathbf{r})\cdot\mathbf{E}(\mathbf{r})\,dv \tag{2.78}$$

We can consider the volume V (enclosed by S) to be as large as we like, say a sphere of radius R, so that all charges are contained well within it. Then the surface integral over S vanishes as $R\to\infty$ since $\phi \sim 1/R$ and $E \sim 1/R^2$, whereas $da \sim R^2$. Then the total energy is expressed solely in terms of the electric field:

$$U = \int\limits_{\substack{\text{All}\\ \text{Space}}} u_E(\mathbf{r})\,dv \tag{2.79}$$

where

$$u_E(\mathbf{r}) \equiv \frac{\epsilon_0 E^2}{2} \tag{2.80}$$

is interpreted as the energy density of the electric field.

Appendix 2.D Energy Density of Magnetic Fields

In a derivation similar to that in appendix 2.C, we consider the energy associated with magnetic fields.[23] The magnetic vector potential $\mathbf{A}(\mathbf{r})$ (in units of energy per charge×velocity) describes the potential energy that a charge moving with velocity $\mathbf{v}$ would experience if placed in the field. The magnetic field and the vector potential are connected through

$$\mathbf{B}(\mathbf{r}) = \nabla \times \mathbf{A}(\mathbf{r}) \tag{2.81}$$

The energy U necessary to assemble a distribution of currents can be written in terms of a summation over all of the currents (or current density $\mathbf{J}(\mathbf{r})$) located within the vector potential field:

$$U = \frac{1}{2}\int_V \mathbf{J}(\mathbf{r}) \cdot \mathbf{A}(\mathbf{r})\, dv \tag{2.82}$$

As in (2.75), the factor 1/2 is necessary to avoid double counting the influence of the currents on each other.

Under the assumption of steady currents (no variations in time), we may substitute Ampere's law (1.21) into (2.82), which yields

$$U = \frac{1}{2\mu_0}\int_V [\nabla \times \mathbf{B}(\mathbf{r})] \cdot \mathbf{A}(\mathbf{r})\, dv \tag{2.83}$$

Next we employ the vector identity P0.8 from which the previous expression becomes

$$U = \frac{1}{2\mu_0}\int_V \mathbf{B}(\mathbf{r}) \cdot [\nabla \times \mathbf{A}(\mathbf{r})]\, dv - \frac{1}{2\mu_0}\int_V \nabla \cdot [\mathbf{A}(\mathbf{r}) \times \mathbf{B}(\mathbf{r})]\, dv \tag{2.84}$$

Upon substituting (2.81) into the first equation and applying the Divergence theorem (0.11) on the second integral, this expression for total energy becomes

$$U = \frac{1}{2\mu_0}\int_V \mathbf{B}(\mathbf{r}) \cdot \mathbf{B}(\mathbf{r})\, dv - \frac{1}{2\mu_0}\oint_S [\mathbf{A}(\mathbf{r}) \times \mathbf{B}(\mathbf{r})] \cdot \hat{\mathbf{n}}\, da \tag{2.85}$$

As was done in connection with (2.78), if we choose a large enough volume (a sphere with radius $R \to \infty$), the surface integral vanishes since $A \sim 1/R$ and $B \sim 1/R^2$, whereas $da \sim R^2$. The total energy (2.85) then reduces to

$$U = \int_{\substack{\text{All}\\ \text{Space}}} u_B(\mathbf{r})\, dv \tag{2.86}$$

where

$$u_B(\mathbf{r}) \equiv \frac{B^2}{2\mu_0} \tag{2.87}$$

is the energy density for a magnetic field.

---

[23] J. R. Reitz, F. J. Milford, and R. W. Christy, *Foundations of Electromagnetic Theory* 3rd ed., Sect. 12-2 (Reading, Massachusetts: Addison-Wesley, 1979.

Exercises

Exercises for 2.4 The Lorentz Model of Dielectrics

P2.1 Verify that (2.35) is a solution to (2.34).

P2.2 Derive the Sellmeier equation

$$n^2 = 1 + \frac{A\lambda_{\text{vac}}^2}{\lambda_{\text{vac}}^2 - \lambda_{0,\text{vac}}^2}$$

from (2.39) for a gas with negligible absorption (i.e. $\gamma \cong 0$, valid far from resonance ω_0), where $\lambda_{0,\text{vac}}$ corresponds to frequency ω_0 and A is a constant. Many materials (e.g. glass, air) have strong resonances in the ultraviolet. In such materials, do you expect the index of refraction for blue light to be greater than that for red light? Make a sketch of n as a function of wavelength for visible light down to the ultraviolet (where $\lambda_{0,\text{vac}}$ is located).

P2.3 In the Lorentz model, take $N = 10^{28}\ \text{m}^{-3}$ for the density of bound electrons in an insulator, and a single transition at $\omega_0 = 6 \times 10^{15}$ rad/sec (in the UV), and damping $\gamma = \omega_0/5$ (quite broad). Assume that the magnitude of $\mathbf{E}_0$ is 10^4 V/m. For three frequencies i) $\omega = \omega_0 - 2\gamma$, ii) $\omega = \omega_0$, and iii) $\omega = \omega_0 + 2\gamma$ find:

(a) the amplitude and phase of the charge displacement $\mathbf{r}_e$ (2.35) relative to the phase of $\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$.

(b) the magnitude and complex phase of the susceptibility $\chi(\omega)$. Does $\chi(\omega)$ depend on the strength of the E-field?

(c) n and κ at the three frequencies via (2.29) and (2.27).

Answer: i) $n = 1.53$, $\kappa = 0.0817$, ii) $n = 1.66$, $\kappa = 1.33$, iii) $n = 0.470$, $\kappa = 0.263$.

(d) the three speeds of light in terms of c and how far light penetrates into the material before only $1/e$ of the amplitude of $\mathbf{E}$ remains.

P2.4 (a) Use a computer to plot n and κ as a function of ω for a dielectric (i.e. obtain graphs such as the ones in Fig. 2.5). Use the Lorentz model and the following parameters: $\omega_0 = 10\omega_\text{p}$, and $\gamma = \omega_\text{p}$; plot your function from $\omega = 0$ to $\omega = 20\omega_\text{p}$. No need to choose a value for ω_p; your horizontal axis will be in units of ω_p.

(b) Plot n and κ as a function of frequency for a material that has three resonant frequencies: $\omega_{01} = 10\omega_\text{p}$, $\gamma_1 = \omega_\text{p}$, $f_1 = 0.5$; $\omega_{02} = 15\omega_\text{p}$, $\gamma_2 = \omega_\text{p}$, $f_2 = 0.25$; and $\omega_{03} = 25\omega_\text{p}$, $\gamma_3 = 3\omega_\text{p}$, $f_3 = 0.25$. Plot the results from $\omega = 0$ to $\omega = 30\omega_\text{p}$.

Exercises for 2.5 Index of Refraction of a Conductor

P2.5 For silver, the complex refractive index is characterized by $n = 0.13$ and $\kappa = 4.0$.[24] Find the distance that light travels inside of silver before the field is reduced by a factor of $1/e$. Assume a wavelength of $\lambda_{\text{vac}} =$ 633 nm. What is the speed of the wave crests in the silver (written as a number times c)? Are you surprised?

P2.6 Use (2.48) and expressions that follow (2.48) to calculate the index of silver at $\lambda = 633$nm. The density of free electrons in silver is $N = 5.86 \times 10^{28}\text{m}^{-3}$ and the DC conductivity is $\sigma = 6.62 \times 10^7 C^2/(\text{J} \cdot \text{m} \cdot \text{s})$.[25] Compare with the actual index given in P2.5.

Answer: $n + i\kappa = 0.02 + i4.5$

P2.7 The uppermost part of the atmosphere is ionized by solar radiation, which creates a low-density plasma called the ionosphere. Note: $\omega_0 = 0$ and $\gamma = 0$.

(a) If the index of refraction of the ionosphere is $\mathcal{N} = 0.9$ for an FM station at $\nu = \omega/2\pi = 100$ MHz, calculate the number of free electrons per cubic meter.

(b) What is the complex refractive index of the ionosphere for an AM radio station at 1160 kHz? Is this frequency above or below the plasma frequency? Assume the same density of free electrons as in part (a).

For your information, AM radio reflects better than FM radio from the ionosphere (like visible light from a metal mirror). At night, the lower layer of the ionosphere goes away so that AM radio waves reflect from a higher layer.

P2.8 Use a computer to plot n and κ as a function of frequency for a conductor (obtain plots such as the ones in Fig. 2.7). Let $\gamma = 0.02\omega_p$, and plot your function from $\omega = 0.6\omega_p$ to $\omega = 2\omega_p$.

Exercises for 2.7 Irradiance of a Plane Wave

P2.9 In the case of a linearly-polarized plane wave, where the phase of each vector component of $\mathbf{E}_0$ is the same, re-derive (2.62) directly from the real field (2.21). For simplicity, you may ignore absorption (i.e. $\kappa \cong 0$).

HINT: The time-average of $\cos^2\left(\mathbf{k}\cdot\mathbf{r} - \omega t + \phi\right)$ is $1/2$.

P2.10 (a) Find the intensity (in W/cm^2) produced by a short laser pulse with duration $\Delta t = 2.5 \times 10^{-14}$ s and energy $E = 100$ mJ, focused in vacuum to a round spot with radius $r = 5\ \mu$m.

---

[24] *Handbook of Optical Constants of Solids*, Edited by E. D. Palik (Elsevier, 1997).

[25] G. Burns, *Solid State Physics*, p. 194 (Orlando: Academic Press, 1985).

(b) What is the peak electric field E_x (assuming $E_y = E_z = 0$) in units of V/Å?

HINT: The SI units of electric field are N/C = V/m.

(c) What is the peak magnetic field (in T = kg/(s·C)?

P2.11 (a) What is the intensity (in $\mathrm{W/cm^2}$) *on the retina* when looking directly at the sun? Assume that the eye's pupil has a radius $r_{\text{pupil}} = 1$ mm. Take the Sun's irradiance at the earth's surface to be 1.1 $\mathrm{kW/m^2}$, and neglect refractive index (i.e. set $n = 1$). HINT: The Earth-Sun distance is $d_o = 1.5 \times 10^8$ km and the pupil-retina distance is $d_i = 22$ mm. The radius of the Sun $r_{\text{Sun}} = 7.0 \times 10^5$ km is de-magnified on the retina according to the ratio d_i/d_o.

(b) What is the intensity at the retina when looking directly into a 1 mW HeNe laser? Assume that the smallest radius of the laser beam is $r_{\text{waist}} = 0.5$ mm positioned $d_o = 2$ m in front of the eye, and that the entire beam enters the pupil. Compare with part (a).

P2.12 Show that the magnetic field of an intense laser with $\lambda = 1\ \mu$m becomes important for a free electron oscillating in the field at intensities above $10^{18}\ \mathrm{W/cm^2}$. This marks the transition to relativistic physics. Nevertheless, for convenience, use classical physics in making the estimate.

HINT: At lower intensities, the oscillating electric field dominates, so the electron motion can be thought of as arising solely from the electric field. Use this motion to calculate the magnetic force on the moving electron, and compare it to the electric force. The forces become comparable at $10^{18}\ \mathrm{W/cm^2}$.

Exercises for 2.A Radiometry, Photometry, and Color

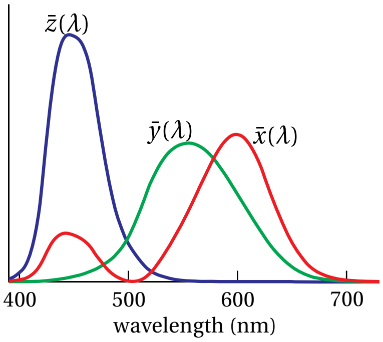


Figure 2.13 Color matching functions for the CIE XYZ color space.

P2.13 The CIE1931 RGB color matching function $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, and $\bar{b}(\lambda)$ can be transformed using (2.65) to obtain color matching functions for the XYZ basis: $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$, plotted in Fig 2.13. As with the RGB color matching functions, the XYZ color matching functions can be used to calculate the color coordinates in the XYZ basis for an arbitrary spectrum:

$$X = \int I(\lambda)\bar{x}d\lambda \qquad Y = \int I(\lambda)\bar{y}d\lambda \qquad Z = \int I(\lambda)\bar{z}d\lambda \tag{2.88}$$

The function $\bar{y}(\lambda)$ was chosen to be exactly the scotopic response curve (shown in Fig. 2.9), so that Y describes the photometric brightness of the light.

(a) Obtain the XYZ color matching functions from www.cvrl.org and calculate the luminous power for a light source with a radiometric spectrum

$$I(\lambda) = I_0 e^{-[(\lambda - 500\text{ nm})/(20\text{ nm})]^2}$$

with $I_0 = 1$ W/nm. HINT: Remember that the response function $\bar{y}$ is the photometric response function of the eye. The standard units from the website give $\bar{y}$ with a peak value of one, so you'll need to use the fact that 1 lm = 1/683 W at the peak of the response curve to get the units right.

(b) Calculate the XYZ color coordinates for the light source in (a).

(c) Calculate the normalized x, y, and z components defined by

$$x = \frac{X}{X+Y+Z}$$

$$y = \frac{Y}{X+Y+Z}$$

$$z = \frac{Z}{X+Y+Z} = 1 - x - y$$

Locate this color on the *chromaticity diagram* in Fig. 2.14. Describe what color light with this spectrum would appear, and how it is possible to represent it using just two coordinates (x and y) as on the diagram. (HINT: You can display a color with bright primaries or dim primaries without changing the color as long as the color of the primaries doesn't change.)

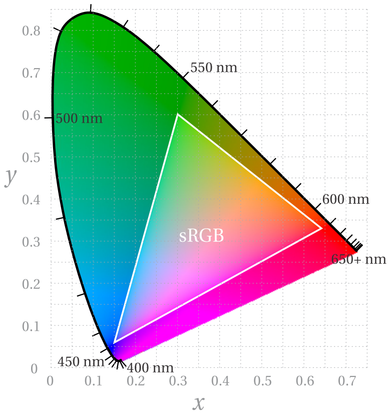


Figure 2.14 A chromaticity diagram plotting the colors visible to the human eye versus x and y, as defined in P2.13. The colors of single-wavelength light lie along the dark line around the edge of the diagram, while the colors that can be displayed by standard computer and TV displays fall inside the white sRGB triangle. This image was created using sRGB encoding so colors outside the sRGB triangle can only be approximated in the chart. All display systems suffer from a limited gamut like this, so the only way to experience the vivid sensation of single-wavelength light is to view the scattered light from a laser (or a gas discharge with a grating) in person.

P2.14 LEDs used in home lighting typically have a power spectrum that looks similar to this function:

$$I(\lambda) = I_0 \left[e^{\left(\frac{\lambda - 460\text{ nm}}{15\text{ nm}}\right)^2} + 0.4 e^{\left(\frac{\lambda - 560\text{ nm}}{90\text{ nm}}\right)^2} \right]$$

where the narrow peak is a blue LED and the broad peak is a yellow phosphor coating that is deposited over the blue LED. Use the process below to display the color of this LED on a computer display.

(a) Obtain a copy of the XYZ color matching functions (available at www.cvrl.org) and calculate the XYZ coordinates using the process described in P2.13.

(b) Now transform the XYZ coordinates into the $(\tilde{R}, \tilde{G}, \tilde{B})$ basis using

$$\begin{bmatrix} \tilde{R} \\ \tilde{G} \\ \tilde{B} \end{bmatrix} = \begin{bmatrix} 3.2406 & -1.5372 & -0.4986 \\ -0.9689 & 1.8758 & 0.0415 \\ 0.0557 & -0.2040 & 1.0570 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

Adjust I_0 so that the largest XYZ coordinate has a numerical value of one.

(c) The $(\tilde{R}, \tilde{G}, \tilde{B})$ values reflect a linear scaling of the stimulus values, but your eyes respond logarithmically, not linearly. This design feature allows you to view faint and bright stars in the night sky at the same time, even though their brightness differs by several orders of magnitude. The sRGB standard approximates the response of the eye by taking each $(\tilde{R}, \tilde{G}, \tilde{B})$ value, represented by $\tilde{C}$, and mapping it to the corresponding component in the sRGB basis like this:

$$C_{sRGB} = \begin{cases} 12.92\tilde{C} & \tilde{C} \le 0.0031308 \\ 1.055\tilde{C}^{\frac{1}{2.4}} - 0.055 & \tilde{C} > 0.0031308 \end{cases}$$

Perform this transformation for each component to find the sRGB components of your light. Then use a computer program (or web site) to display the color of this LED light. It is common to scale your sRGB values to a maximum of 255 for use with 24-bit color rendering (8 bits per channel).

Chapter 3

Reflection and Refraction

As we know from everyday experience, when light arrives at an interface between materials it partially reflects and partially transmits. In this chapter, we examine what happens to plane waves when they propagate from one material (characterized by indices n or even by complex index $\mathcal{N}$) to another material. We will derive expressions to quantify the amount of reflection and transmission. The results depend on the angle of incidence (i.e. the angle between $\mathbf{k}$ and the surface normal) as well as on the orientation of the electric field (called polarization – not to be confused with $\mathbf{P}$, also called polarization). In this chapter, we consider only isotropic materials (e.g. glass); in chapter 5 we consider anisotropic materials (e.g. a crystal).

As we develop the connection between incident, reflected, and transmitted light waves,[1] several familiar relationships will emerge naturally (e.g. Snell's law and Brewster's angle). The formalism also describes polarization-dependent phase shifts upon reflection (especially interesting in the case of reflections from metals).

For simplicity, we initially neglect the imaginary part of the refractive index. Each plane wave is thus characterized by a real wave vector $\mathbf{k}$. We will write each plane wave in the form $\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 \exp\left[i\left(\mathbf{k}\cdot\mathbf{r} - \omega t\right)\right]$, where, as usual, only the real part of the field corresponds to the physical field. The restriction to real refractive indices is not as serious as it might seem. The use of the letter n instead of $\mathcal{N}$ hardly matters. The math is all the same, which demonstrates the power of the complex notation. We can simply update our expressions in the end to include complex refractive indices, but in the meantime it is easier to think of absorption as negligible.

3.1 Refraction at an Interface

Consider a planar boundary between two materials with different indices. Let index n_i characterize the material on the left, and the index n_t characterize the material on the right, as depicted in the Fig. 3.1. When a plane wave traveling in the direction $\mathbf{k}_\mathrm{i}$ is incident on the boundary from the left, it gives rise to a reflected plane wave traveling in the direction $\mathbf{k}_\mathrm{r}$ and a transmitted plane wave traveling in the direction $\mathbf{k}_\mathrm{t}$. The incident and reflected waves exist only to the left of the material interface, and the transmitted wave exists only to the right of the interface. The angles θ_i, θ_r, and θ_t give the angles that each respective wave vector ($\mathbf{k}_\mathrm{i}$, $\mathbf{k}_\mathrm{r}$, and $\mathbf{k}_\mathrm{t}$) makes with the normal to the interface.

---

[1] See M. Born and E. Wolf, *Principles of Optics*, 7th ed., Sect. 1.5 (Cambridge University Press, 1999).

For simplicity, we'll assume that both of the materials are isotropic here. (Chapter 5 discusses refraction for anisotropic materials.) In this case, $\mathbf{k}_\mathrm{i}$, $\mathbf{k}_\mathrm{r}$, and $\mathbf{k}_\mathrm{t}$ all lie in a single plane, referred to as the *plane of incidence*, (i.e. the plane represented by the surface of this page). We are free to orient our coordinate system in many different ways (and every textbook seems to do it differently!).[2] We choose the y–z plane to be the plane of incidence, with the z-direction normal to the interface and the x-axis pointing into the page.

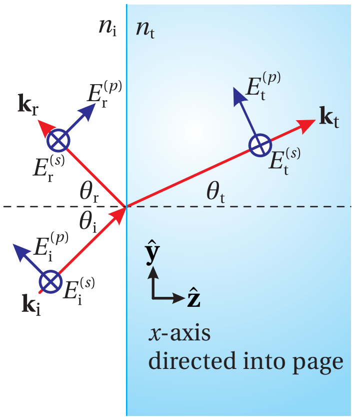


Figure 3.1 Incident, reflected, and transmitted plane wave fields at a material interface.

The electric field vector for each plane wave is confined to a plane perpendicular to its wave vector. We are free to decompose the field vector into arbitrary components as long as they are perpendicular to the wave vector. It is customary to choose one of the electric field vector components to be that which lies within the plane of incidence. We call this *p-polarized light*, where *p* stands for *parallel* to the plane of incidence. The remaining electric field vector component is directed normal to the plane of incidence and is called *s-polarized light*. The *s* stands for *senkrecht*, a German word meaning perpendicular.

Using this system, we can decompose the electric field vector $\mathbf{E}_\mathrm{i}$ into its p-polarized component $E_\mathrm{i}^{(p)}$ and its s-polarized component $E_\mathrm{i}^{(s)}$, as depicted in Fig. 3.1. The s component $E_\mathrm{i}^{(s)}$ is represented by the tail of an arrow pointing into the page, or the x-direction in our convention. The other fields $\mathbf{E}_\mathrm{r}$ and $\mathbf{E}_\mathrm{t}$ are similarly split into s and p components as indicated in Fig. 3.1. All field components are considered to be positive when they point in the direction of their respective arrows.[3] Note that the s-polarized components are parallel for all three plane waves, whereas the p-polarized components are not (except at normal incidence) because each plane wave travels in a different direction.

By inspection of Fig. 3.1, we can write the various wave vectors in terms of the $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ unit vectors:

$$\begin{aligned}
\mathbf{k}_\mathrm{i} &= k_\mathrm{i}\left(\hat{\mathbf{y}}\sin\theta_\mathrm{i} + \hat{\mathbf{z}}\cos\theta_\mathrm{i}\right)\\
\mathbf{k}_\mathrm{r} &= k_\mathrm{r}\left(\hat{\mathbf{y}}\sin\theta_\mathrm{r} - \hat{\mathbf{z}}\cos\theta_\mathrm{r}\right)\\
\mathbf{k}_\mathrm{t} &= k_\mathrm{t}\left(\hat{\mathbf{y}}\sin\theta_\mathrm{t} + \hat{\mathbf{z}}\cos\theta_\mathrm{t}\right)
\end{aligned} \tag{3.1}$$

Also by inspection of Fig. 3.1 (following the conventions for the electric fields indicated by the arrows), we can write the incident, reflected, and transmitted fields in terms of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$:

---

[2] For example, our convention is different than that used by E. Hecht, *Optics*, 3rd ed., Sect. 4.6.2 (Massachusetts: Addison-Wesley, 1998).

[3] Many textbooks draw the arrow for $E_\mathrm{r}^{(p)}$ in the direction opposite of ours. However, that choice leads to an awkward situation at normal incidence (i.e. $\theta_\mathrm{i} = \theta_\mathrm{r} = 0$) where the arrows for the incident and reflected fields are parallel for the s-component but anti parallel for the p-component.

$$\begin{aligned}
\mathbf{E}_\mathrm{i} &= \left[E_\mathrm{i}^{(p)}\left(\hat{\mathbf{y}}\cos\theta_\mathrm{i}-\hat{\mathbf{z}}\sin\theta_\mathrm{i}\right)+\hat{\mathbf{x}}E_\mathrm{i}^{(s)}\right]e^{i\left[k_\mathrm{i}\left(y\sin\theta_\mathrm{i}+z\cos\theta_\mathrm{i}\right)-\omega_\mathrm{i}t\right]}\\
\mathbf{E}_\mathrm{r} &= \left[E_\mathrm{r}^{(p)}\left(\hat{\mathbf{y}}\cos\theta_\mathrm{r}+\hat{\mathbf{z}}\sin\theta_\mathrm{r}\right)+\hat{\mathbf{x}}E_\mathrm{r}^{(s)}\right]e^{i\left[k_\mathrm{r}\left(y\sin\theta_\mathrm{r}-z\cos\theta_\mathrm{r}\right)-\omega_\mathrm{r}t\right]}\\
\mathbf{E}_\mathrm{t} &= \left[E_\mathrm{t}^{(p)}\left(\hat{\mathbf{y}}\cos\theta_\mathrm{t}-\hat{\mathbf{z}}\sin\theta_\mathrm{t}\right)+\hat{\mathbf{x}}E_\mathrm{t}^{(s)}\right]e^{i\left[k_\mathrm{t}\left(y\sin\theta_\mathrm{t}+z\cos\theta_\mathrm{t}\right)-\omega_\mathrm{t}t\right]}
\end{aligned}\tag{3.2}$$

Each field has the form (2.8). We have utilized the k-vectors (3.1) in the exponents of (3.2).

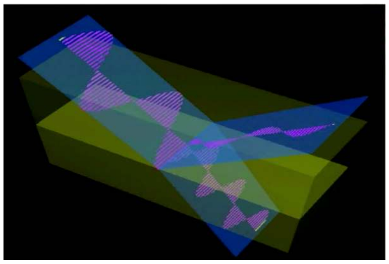


Figure 3.2 Animation of s- and p-polarized fields incident on an interface as the angle of incidence is varied.

Now we are ready to connect the fields on one side of the interface to the fields on the other side. This is done using *boundary conditions*. As explained in appendix 3.A, Maxwell's equations require the components of $\mathbf{E}$ that are parallel to the interface to be the same on either side of the boundary. In our coordinate system, the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ components are parallel to the interface, whereas $z = 0$ defines the interface. This means that at $z = 0$ the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ components of the combined incident and reflected fields must equal the corresponding components of the transmitted field:

$$\begin{aligned}
\left[E_\mathrm{i}^{(p)}\hat{\mathbf{y}}\cos\theta_\mathrm{i}+\hat{\mathbf{x}}E_\mathrm{i}^{(s)}\right]e^{i\left(k_\mathrm{i}y\sin\theta_\mathrm{i}-\omega_\mathrm{i}t\right)}&+\left[E_\mathrm{r}^{(p)}\hat{\mathbf{y}}\cos\theta_\mathrm{r}+\hat{\mathbf{x}}E_\mathrm{r}^{(s)}\right]e^{i\left(k_\mathrm{r}y\sin\theta_\mathrm{r}-\omega_\mathrm{r}t\right)}\\
&=\left[E_\mathrm{t}^{(p)}\hat{\mathbf{y}}\cos\theta_\mathrm{t}+\hat{\mathbf{x}}E_\mathrm{t}^{(s)}\right]e^{i\left(k_\mathrm{t}y\sin\theta_\mathrm{t}-\omega_\mathrm{t}t\right)}
\end{aligned}\tag{3.3}$$

Since this equation must hold for all conceivable values of t and y, we are compelled to set all the phase factors in the complex exponentials equal to each other. The time portion of the phase factors requires the frequency of all waves to be the same:

$$\omega_\mathrm{i} = \omega_\mathrm{r} = \omega_\mathrm{t} \equiv \omega \tag{3.4}$$

(We could have guessed that all frequencies would be the same; otherwise wavefronts would be annihilated or created at the interface.) Similarly, equating the spatial terms in the exponents of (3.3) requires

$$k_\mathrm{i}\sin\theta_\mathrm{i} = k_\mathrm{r}\sin\theta_\mathrm{r} = k_\mathrm{t}\sin\theta_\mathrm{t} \tag{3.5}$$

Now recall from (2.19) the relations $k_\mathrm{i} = k_\mathrm{r} = n_\mathrm{i}\omega/c$ and $k_\mathrm{t} = n_\mathrm{t}\omega/c$. With these relations, (3.5) yields the *law of reflection*

$$\theta_\mathrm{r} = \theta_\mathrm{i} \tag{3.6}$$

and *Snell's law*

$$n_\mathrm{i}\sin\theta_\mathrm{i} = n_\mathrm{t}\sin\theta_\mathrm{t} \tag{3.7}$$

The three angles θ_i, θ_r, and θ_t are not independent. The reflected angle matches the incident angle, and the transmitted angle obeys Snell's law. The phenomenon of *refraction* refers to the fact that θ_i and θ_t are different. That is, light 'bends' as it transmits through an interface.

Willebrord Snell (or Snellius) (1580–1626, Dutch) was an astronomer and mathematician born in Leiden, Netherlands. In 1613 he succeeded his father as professor of mathematics at the University of Leiden. He was an accomplished mathematician, developing a new method for calculating π as well as an improved method for measuring the circumference of the earth. He is most famous for his rediscovery of the law of refraction in 1621. (The law was known (in table form) to the ancient Greek mathematician Ptolemy, to Persian engineer Ibn Sahl (900s), and to Polish philosopher Witelo (1200s).) Snell authored several books, including one on trigonometry, published a year after his death. (Wikipedia)

Because the exponents are all identical, (3.3) reduces to two relatively simple equations (one for each dimension, $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$):

$$E_{\mathrm{i}}^{(s)} + E_{\mathrm{r}}^{(s)} = E_{\mathrm{t}}^{(s)} \tag{3.8}$$

and

$$\left(E_{\mathrm{i}}^{(p)} + E_{\mathrm{r}}^{(p)}\right)\cos\theta_{\mathrm{i}} = E_{\mathrm{t}}^{(p)}\cos\theta_{\mathrm{t}} \tag{3.9}$$

We have derived these equations from the boundary condition (3.54) on the parallel component of the electric field. This set of equations has four unknowns ($E_{\mathrm{r}}^{(p)}$, $E_{\mathrm{r}}^{(s)}$, $E_{\mathrm{t}}^{(p)}$, and $E_{\mathrm{t}}^{(s)}$), assuming that we pick the incident fields. We require two additional equations to solve the system. These are obtained using the separate boundary condition on the parallel component of magnetic fields given in (3.58) (also discussed in appendix 3.A).

From Faraday's law (1.3), we have for a plane wave (see (2.56))

$$\mathbf{B} = \frac{\mathbf{k}\times\mathbf{E}}{\omega} = \frac{n}{c}\hat{\mathbf{u}}\times\mathbf{E} \tag{3.10}$$

where $\hat{\mathbf{u}} \equiv \mathbf{k}/k$ is a unit vector in the direction of $\mathbf{k}$. We have also utilized (2.19) for a real index. This expression is useful for writing $\mathbf{B}_{\mathrm{i}}$, $\mathbf{B}_{\mathrm{r}}$, and $\mathbf{B}_{\mathrm{t}}$ in terms of the electric field components that we have already introduced. When injecting (3.1) and (3.2) into (3.10), the incident, reflected, and transmitted magnetic fields turn out to be

$$\begin{aligned}
\mathbf{B}_{\mathrm{i}} &= \frac{n_{\mathrm{i}}}{c}\left[-\hat{\mathbf{x}}E_{\mathrm{i}}^{(p)} + E_{\mathrm{i}}^{(s)}\left(-\hat{\mathbf{z}}\sin\theta_{\mathrm{i}} + \hat{\mathbf{y}}\cos\theta_{\mathrm{i}}\right)\right]e^{i[k_{\mathrm{i}}(y\sin\theta_{\mathrm{i}}+z\cos\theta_{\mathrm{i}})-\omega_{\mathrm{i}}t]}\\
\mathbf{B}_{\mathrm{r}} &= \frac{n_{\mathrm{r}}}{c}\left[\hat{\mathbf{x}}E_{\mathrm{r}}^{(p)} + E_{\mathrm{r}}^{(s)}\left(-\hat{\mathbf{z}}\sin\theta_{\mathrm{r}} - \hat{\mathbf{y}}\cos\theta_{\mathrm{r}}\right)\right]e^{i[k_{\mathrm{r}}(y\sin\theta_{\mathrm{r}}-z\cos\theta_{\mathrm{r}})-\omega_{\mathrm{r}}t]}\\
\mathbf{B}_{\mathrm{t}} &= \frac{n_{\mathrm{t}}}{c}\left[-\hat{\mathbf{x}}E_{\mathrm{t}}^{(p)} + E_{\mathrm{t}}^{(s)}\left(-\hat{\mathbf{z}}\sin\theta_{\mathrm{t}} + \hat{\mathbf{y}}\cos\theta_{\mathrm{t}}\right)\right]e^{i[k_{\mathrm{t}}(y\sin\theta_{\mathrm{t}}+z\cos\theta_{\mathrm{t}})-\omega_{\mathrm{t}}t]}
\end{aligned} \tag{3.11}$$

Next, we apply the boundary condition (3.58), namely that the components of $\mathbf{B}$ parallel to the interface (i.e. in the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ dimensions) are the same[4] on either side of the plane $z = 0$. Since we already know that the exponents are all equal and that $\theta_{\mathrm{r}} = \theta_{\mathrm{i}}$ and $n_{\mathrm{i}} = n_{\mathrm{r}}$, the boundary condition gives

$$\frac{n_{\mathrm{i}}}{c}\left[-\hat{\mathbf{x}}E_{\mathrm{i}}^{(p)} + E_{\mathrm{i}}^{(s)}\hat{\mathbf{y}}\cos\theta_{\mathrm{i}}\right] + \frac{n_{\mathrm{i}}}{c}\left[\hat{\mathbf{x}}E_{\mathrm{r}}^{(p)} - E_{\mathrm{r}}^{(s)}\hat{\mathbf{y}}\cos\theta_{\mathrm{i}}\right] = \frac{n_{\mathrm{t}}}{c}\left[-\hat{\mathbf{x}}E_{\mathrm{t}}^{(p)} + E_{\mathrm{t}}^{(s)}\hat{\mathbf{y}}\cos\theta_{\mathrm{t}}\right] \tag{3.12}$$

As before, (3.12) reduces to two relatively simple equations (one for the $\hat{\mathbf{x}}$ dimension and one for the $\hat{\mathbf{y}}$ dimension):

$$n_{\mathrm{i}}\left(E_{\mathrm{i}}^{(p)} - E_{\mathrm{r}}^{(p)}\right) = n_{\mathrm{t}}E_{\mathrm{t}}^{(p)} \tag{3.13}$$

and

$$n_{\mathrm{i}}\left(E_{\mathrm{i}}^{(s)} - E_{\mathrm{r}}^{(s)}\right)\cos\theta_{\mathrm{i}} = n_{\mathrm{t}}E_{\mathrm{t}}^{(s)}\cos\theta_{\mathrm{t}} \tag{3.14}$$

These two equations together with (3.8) and (3.9) allow us to solve for the reflected $\mathbf{E}_{\mathrm{r}}$ and transmitted fields $\mathbf{E}_{\mathrm{t}}$ for the s and p polarization components. However, (3.8), (3.9), (3.13), and (3.14) are not yet in their most convenient form.

---

[4] We assume the permeability μ is the same everywhere—no magnetic effects.

3.2 The Fresnel Coefficients

Augustin Fresnel first derived the equations in the previous section. Since he lived well before Maxwell's time, he did not have the benefit of Maxwell's equations as we have. Instead, Fresnel thought of light as transverse mechanical waves propagating within materials. (Fresnel was naturally a proponent of luminiferous ether.) Instead of relating the parallel components of the electric and magnetic fields across the boundary between the materials, Fresnel used the principle that the two materials should not slip relative to each other at the boundary. This 'gluing' of the materials at the interface also forbids the possibility of gaps or the like forming at the interface as the two materials experience wave vibrations. This mechanical approach to light worked splendidly, arriving at the same results that we obtained from our modern viewpoint.

Fresnel wrote the relationships between the various plane waves depicted in Fig. 3.1 in terms of coefficients that compare the reflected and transmitted field amplitudes to those of the incident field. In the following example, we illustrate this procedure for s-polarized light. It is left as a homework exercise to solve the equations for p-polarized light (see P3.1).

Augustin Fresnel (1788–1829, French) was born in Broglie, France, the son of an architect. As a child, he was slow to develop and still could not read when he was eight years old, but by age sixteen he excelled and entered the École Polytechnique where he earned distinction. As a young man, Fresnel began a successful career as an engineer, but he lost his post in 1814 when Napoleon returned to power. (Fresnel had supported the Bourbons.) This difficult year was when Fresnel turned his attention to optics. Fresnel became a major proponent of the wave theory of light and four years later wrote a paper on diffraction for which he was awarded a prize by the French Academy of Sciences. A year later he was appointed commissioner of lighthouses, which motivated the invention of the Fresnel lens (still used in many commercial applications). Fresnel was underappreciated before his untimely death from tuberculosis. Many of his papers did not make it into print until years later. Fresnel made huge advances in the understanding of reflection, diffraction, polarization, and birefringence. In 1824 Fresnel wrote to Thomas Young, "All the compliments that I have received from Arago, Laplace and Biot never gave me so much pleasure as the discovery of a theoretic truth, or the confirmation of a calculation by experiment." Augustin Fresnel is a hero of one of the authors of this textbook. (Wikipedia)

Example 3.1

Calculate the ratio of transmitted field to the incident field and the ratio of the reflected field to incident field for s-polarized light.

Solution: We write (3.8) and (3.14) as

$$E_\mathrm{i}^{(s)} + E_\mathrm{r}^{(s)} = E_\mathrm{t}^{(s)} \qquad \text{and} \qquad E_\mathrm{i}^{(s)} - E_\mathrm{r}^{(s)} = \frac{n_\mathrm{t}\cos\theta_\mathrm{t}}{n_\mathrm{i}\cos\theta_\mathrm{i}} E_\mathrm{t}^{(s)} \tag{3.15}$$

Adding these two equations yields

$$2E_\mathrm{i}^{(s)} = \left[1 + \frac{n_\mathrm{t}\cos\theta_\mathrm{t}}{n_\mathrm{i}\cos\theta_\mathrm{i}}\right] E_\mathrm{t}^{(s)} \tag{3.16}$$

After a little rearrangement we get

$$\frac{E_\mathrm{t}^{(s)}}{E_\mathrm{i}^{(s)}} = \frac{2n_\mathrm{i}\cos\theta_\mathrm{i}}{n_\mathrm{i}\cos\theta_\mathrm{i} + n_\mathrm{t}\cos\theta_\mathrm{t}} \tag{3.17}$$

To get the ratio of reflected field to incident field, we subtract the equations in (3.15) to get

$$2E_\mathrm{r}^{(s)} = \left[1 - \frac{n_\mathrm{t}\cos\theta_\mathrm{t}}{n_\mathrm{i}\cos\theta_\mathrm{i}}\right] E_\mathrm{t}^{(s)} \tag{3.18}$$

We divide (3.18) by (3.16), and after simplification arrive at

$$\frac{E_\mathrm{r}^{(s)}}{E_\mathrm{i}^{(s)}} = \frac{n_\mathrm{i}\cos\theta_\mathrm{i} - n_\mathrm{t}\cos\theta_\mathrm{t}}{n_\mathrm{i}\cos\theta_\mathrm{i} + n_\mathrm{t}\cos\theta_\mathrm{t}} \tag{3.19}$$

The ratio of the reflected and transmitted field components to the incident field components are specified by the *Fresnel coefficients*, which are defined as follows:

$$r_s \equiv \frac{E_{\mathrm{r}}^{(s)}}{E_{\mathrm{i}}^{(s)}} = \frac{n_{\mathrm{i}}\cos\theta_{\mathrm{i}} - n_{\mathrm{t}}\cos\theta_{\mathrm{t}}}{n_{\mathrm{i}}\cos\theta_{\mathrm{i}} + n_{\mathrm{t}}\cos\theta_{\mathrm{t}}} = \frac{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}} - \sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{t}}}{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}} + \sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{t}}} = \frac{\sin(\theta_{\mathrm{t}} - \theta_{\mathrm{i}})}{\sin(\theta_{\mathrm{t}} + \theta_{\mathrm{i}})} \tag{3.20}$$

$$t_s \equiv \frac{E_{\mathrm{t}}^{(s)}}{E_{\mathrm{i}}^{(s)}} = \frac{2n_{\mathrm{i}}\cos\theta_{\mathrm{i}}}{n_{\mathrm{i}}\cos\theta_{\mathrm{i}} + n_{\mathrm{t}}\cos\theta_{\mathrm{t}}} = \frac{2\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}}}{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}} + \sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{t}}} = \frac{2\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}}}{\sin(\theta_{\mathrm{t}} + \theta_{\mathrm{i}})} \tag{3.21}$$

$$r_p \equiv \frac{E_{\mathrm{r}}^{(p)}}{E_{\mathrm{i}}^{(p)}} = \frac{n_{\mathrm{i}}\cos\theta_{\mathrm{t}} - n_{\mathrm{t}}\cos\theta_{\mathrm{i}}}{n_{\mathrm{i}}\cos\theta_{\mathrm{t}} + n_{\mathrm{t}}\cos\theta_{\mathrm{i}}} = \frac{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{t}} - \sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{i}}}{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{t}} + \sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{i}}} = \frac{\tan(\theta_{\mathrm{t}} - \theta_{\mathrm{i}})}{\tan(\theta_{\mathrm{t}} + \theta_{\mathrm{i}})} \tag{3.22}$$

$$t_p \equiv \frac{E_{\mathrm{t}}^{(p)}}{E_{\mathrm{i}}^{(p)}} = \frac{2n_{\mathrm{i}}\cos\theta_{\mathrm{i}}}{n_{\mathrm{i}}\cos\theta_{\mathrm{t}} + n_{\mathrm{t}}\cos\theta_{\mathrm{i}}} = \frac{2\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}}}{\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{t}} + \sin\theta_{\mathrm{i}}\cos\theta_{\mathrm{i}}} = \frac{2\sin\theta_{\mathrm{t}}\cos\theta_{\mathrm{i}}}{\sin(\theta_{\mathrm{t}} + \theta_{\mathrm{i}})\cos(\theta_{\mathrm{t}} - \theta_{\mathrm{i}})} \tag{3.23}$$

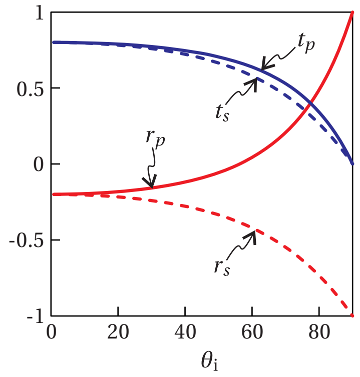


Figure 3.3 The Fresnel coefficients plotted versus θ_{i} for the case of an air-glass interface with $n_{\mathrm{i}} = 1$ and $n_{\mathrm{t}} = 1.5$.

All of the above forms of the Fresnel coefficients are potentially useful, depending on the problem at hand. Remember that the angles in the coefficient are not independently chosen, but are subject to Snell's law (3.7). Snell's law has been used to produce the alternative expressions from the first.

The Fresnel coefficients pin down the electric field amplitudes on the two sides of the boundary. They also keep track of phase shifts at a boundary. In Fig. 3.3 we have plotted the Fresnel coefficients for the case of an air-glass interface. Notice that the reflection coefficients are sometimes negative in this plot, which corresponds to a phase shift of π upon reflection (note $e^{i\pi} = -1$). Later we will see that when absorbing materials are encountered, more complicated phase shifts can arise due to the complex index of refraction.

3.3 Reflectance and Transmittance

We often want to know the fraction of power that reflects from or transmits through an interface. Energy conservation requires the incident power to balance the reflected and transmitted power:

$$P_{\mathrm{i}} = P_{\mathrm{r}} + P_{\mathrm{t}} \tag{3.24}$$

Moreover, the power separates cleanly into power associated with s- and p-polarized fields:

$$P_{\mathrm{i}}^{(s)} = P_{\mathrm{r}}^{(s)} + P_{\mathrm{t}}^{(s)} \qquad \text{and} \qquad P_{\mathrm{i}}^{(p)} = P_{\mathrm{r}}^{(p)} + P_{\mathrm{t}}^{(p)} \tag{3.25}$$

Since power is proportional to intensity (i.e. power per area) and intensity is proportional to the square of the field amplitude. We can write the fraction of reflected power, called *reflectance*, in terms of our previously defined Fresnel coefficients:

$$R_s \equiv \frac{P_{\rm r}^{(s)}}{P_{\rm i}^{(s)}} = \frac{I_{\rm r}^{(s)}}{I_{\rm i}^{(s)}} = \frac{\left|E_{\rm r}^{(s)}\right|^2}{\left|E_{\rm i}^{(s)}\right|^2} = |r_s|^2 \qquad \text{and} \qquad R_p \equiv \frac{P_{\rm r}^{(p)}}{P_{\rm i}^{(p)}} = \frac{I_{\rm r}^{(p)}}{I_{\rm i}^{(p)}} = \frac{\left|E_{\rm r}^{(p)}\right|^2}{\left|E_{\rm i}^{(p)}\right|^2} = \left|r_p\right|^2 \tag{3.26}$$

The total reflected intensity is therefore

$$I_{\rm r} = I_{\rm r}^{(s)} + I_{\rm r}^{(p)} = R_s I_{\rm i}^{(s)} + R_p I_{\rm i}^{(p)} \tag{3.27}$$

where, according to (2.62), the total incident intensity is given by

$$I_{\rm i} = I_{\rm i}^{(s)} + I_{\rm i}^{(p)} = \frac{1}{2} n_{\rm i} \epsilon_0 c \left[\left|E_{\rm i}^{(s)}\right|^2 + \left|E_{\rm i}^{(p)}\right|^2\right] \tag{3.28}$$

From (3.25) and (3.26), the transmitted power is

$$P_{\rm t}^{\rm (s)} = P_{\rm i}^{\rm (s)} - P_{\rm r}^{\rm (s)} = (1 - R_s)\, P_{\rm i}^{\rm (s)} \qquad \text{and} \qquad P_{\rm t}^{\rm (p)} = P_{\rm i}^{\rm (p)} + P_{\rm r}^{\rm (p)} = \left(1 - R_p\right) P_{\rm i}^{\rm (p)} \tag{3.29}$$

From this expression we see that the fraction of the power that transmits, called the *transmittance*, is

$$T_s \equiv \frac{P_{\rm t}^{(s)}}{P_{\rm i}^{(s)}} = 1 - R_s \qquad \text{and} \qquad T_p \equiv \frac{P_{\rm t}^{(p)}}{P_{\rm i}^{(p)}} = 1 - R_p \tag{3.30}$$

Figure 3.4 shows typical reflectance and transmittance values for an air-glass interface.

You might be surprised at first to learn that

$$T_s \neq |t_s|^2 \qquad \text{and} \qquad T_p \neq \left|t_p\right|^2 \tag{3.31}$$

However, recall that the transmitted intensity (in terms of the transmitted fields) depends also on the refractive index. The Fresnel coefficients t_s and t_p relate the bare electric fields to each other, whereas the transmitted intensity is

$$I_{\rm t} = I_{\rm t}^{(s)} + I_{\rm t}^{(p)} = \frac{1}{2} n_{\rm t} \epsilon_0 c \left[\left|E_{\rm t}^{(s)}\right|^2 + \left|E_{\rm t}^{(p)}\right|^2\right] \tag{3.32}$$

In view of (3.28) and (3.32), we expect T_s and T_p to depend on the ratio of the refractive indices $n_{\rm t}$ and $n_{\rm i}$ in addition to $|t_s|^2$ or $\left|t_p\right|^2$.

There is another more subtle reason for the inequalities in (3.31). Consider a lateral strip of light associated with a plane wave incident upon the material interface in Fig. 3.5. Upon refraction into the second medium, the strip is seen to change its width by the factor $\cos\theta_{\rm t}/\cos\theta_{\rm i}$. This is a purely geometrical effect, owing to the change in propagation direction at the interface. Since power is intensity times area, the transmittance picks up this geometrical factor via the ratio of the areas $A_{\rm t}/A_{\rm i}$ as follows:

$$\begin{aligned} T_s &\equiv \frac{P_{\rm t}^{(s)}}{P_{\rm i}^{(s)}} = \frac{I_{\rm t}^{(s)} A_{\rm t}}{I_{\rm i}^{(s)} A_{\rm i}} = \frac{n_{\rm t}\cos\theta_{\rm t}}{n_{\rm i}\cos\theta_{\rm i}} |t_s|^2 \\ T_p &\equiv \frac{P_{\rm t}^{(p)}}{P_{\rm i}^{(p)}} = \frac{I_{\rm t}^{(p)} A_{\rm t}}{I_{\rm i}^{(p)} A_{\rm i}} = \frac{n_{\rm t}\cos\theta_{\rm t}}{n_{\rm i}\cos\theta_{\rm i}} \left|t_p\right|^2 \end{aligned} \qquad \text{(not valid if total internal reflection)} \tag{3.33}$$

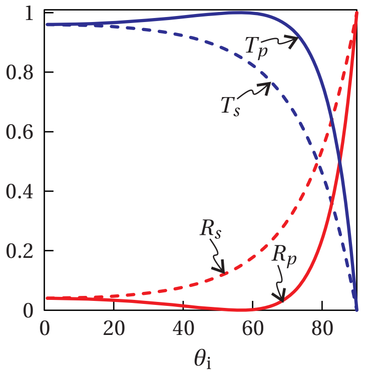


Figure 3.4 The reflectance and transmittance plotted versus $\theta_{\rm i}$ for the case of an air-glass interface with $n_{\rm i} = 1$ and $n_{\rm t} = 1.5$.

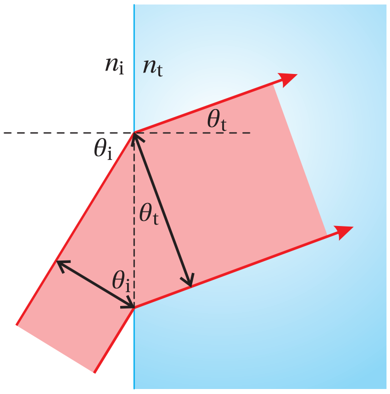


Figure 3.5 Light refracting into a surface

David Brewster (1781–1868, Scottish) was born in Jedburgh, Scotland. His father was a teacher and wanted David to become a clergyman. At age twelve, David went to the University of Edinburgh for that purpose, but his inclination for natural science soon became apparent. He became licensed to preach, but his interests in science distracted him from that profession, and he spent much of his time studying diffraction. Taking an empirical approach, Brewster independently discovered many of the same things usually credited to Fresnel. He even made a dioptric apparatus for lighthouses before Fresnel developed his. Brewster became somewhat famous in his day for the development of the kaleidoscope and stereoscope for enjoyment by the general public. Brewster was a prolific science writer and editor throughout his life. Among his works is an important biography of Isaac Newton. He was knighted for his accomplishments in 1831. (Wikipedia)

Note that (3.33) is valid only if a real angle θ_t exists; it does not hold when the incident angle exceeds the critical angle for total internal reflection, discussed in section 3.5. In that situation, we must stick with (3.30).

Example 3.2

Show analytically that $R_p + T_\mathrm{p} = 1$, where R_p is given by (3.26) and T_p is given by (3.33).

Solution: From (3.22) we have

$$\begin{aligned} R_p &= \left|\frac{n_\mathrm{i}\cos\theta_\mathrm{t} - n_\mathrm{t}\cos\theta_\mathrm{i}}{n_\mathrm{i}\cos\theta_\mathrm{t} + n_\mathrm{t}\cos\theta_\mathrm{i}}\right|^2 \\ &= \frac{n_\mathrm{i}^2\cos^2\theta_\mathrm{t} - 2n_\mathrm{i}n_\mathrm{t}\cos\theta_\mathrm{i}\cos\theta_\mathrm{t} + n_\mathrm{i}^2\cos^2\theta_\mathrm{i}}{(n_\mathrm{i}\cos\theta_\mathrm{t} + n_\mathrm{t}\cos\theta_\mathrm{i})^2} \end{aligned}$$

From (3.23) and (3.33) we have

$$\begin{aligned} T_p &= \frac{n_\mathrm{t}\cos\theta_\mathrm{t}}{n_\mathrm{i}\cos\theta_\mathrm{i}}\left|\frac{2n_\mathrm{i}\cos\theta_\mathrm{i}}{n_\mathrm{i}\cos\theta_\mathrm{t} + n_\mathrm{t}\cos\theta_\mathrm{i}}\right|^2 \\ &= \frac{4n_\mathrm{i}n_\mathrm{t}\cos\theta_\mathrm{i}\cos\theta_\mathrm{t}}{(n_\mathrm{i}\cos\theta_\mathrm{t} + n_\mathrm{t}\cos\theta_\mathrm{i})^2} \end{aligned}$$

Then

$$\begin{aligned} R_p + T_p &= \frac{n_\mathrm{i}^2\cos^2\theta_\mathrm{t} + 2n_\mathrm{i}n_\mathrm{t}\cos\theta_\mathrm{i}\cos\theta_\mathrm{t} + n_\mathrm{i}^2\cos^2\theta_\mathrm{i}}{(n_\mathrm{i}\cos\theta_\mathrm{t} + n_\mathrm{t}\cos\theta_\mathrm{i})^2} \\ &= \frac{(n_\mathrm{i}\cos\theta_\mathrm{t} + n_\mathrm{t}\cos\theta_\mathrm{i})^2}{(n_\mathrm{i}\cos\theta_\mathrm{t} + n_\mathrm{t}\cos\theta_\mathrm{i})^2} = 1 \end{aligned}$$

3.4 Brewster's Angle

Notice r_p and R_p go to zero at a certain angle in Figs. 3.3 and 3.4, indicating that no p-polarized light is reflected at this angle. This behavior is quite general, as we can see from the final form of the Fresnel coefficient formula for r_p in (3.22), which has $\tan(\theta_\mathrm{i} + \theta_\mathrm{t})$ in the denominator. Since the tangent 'blows up' at $\pi/2$, the reflection coefficient goes to zero when

$$\theta_\mathrm{i} + \theta_\mathrm{t} = \frac{\pi}{2} \qquad \text{(requirement for zero } p\text{-polarized reflection)} \qquad (3.34)$$

By inspecting Fig. 3.1, we see that this condition occurs when the reflected and transmitted wave vectors, $\mathbf{k}_\mathrm{r}$ and $\mathbf{k}_\mathrm{t}$, are perpendicular to each other (see also Fig. 3.6). If we insert (3.34) into Snell's law (3.7), we can solve for the incident angle θ_i that gives rise to this special circumstance:

$$n_\mathrm{i}\sin\theta_\mathrm{i} = n_\mathrm{t}\sin\left(\frac{\pi}{2} - \theta_\mathrm{i}\right) = n_\mathrm{t}\cos\theta_\mathrm{i} \qquad (3.35)$$

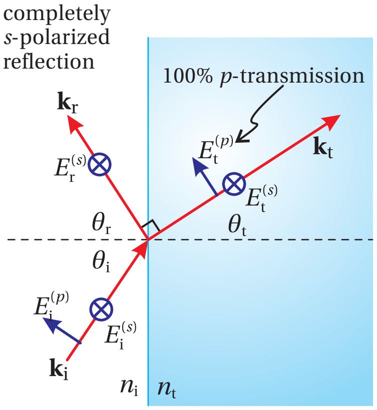


Figure 3.6 Brewster's angle coincides with the situation where $\mathbf{k}_\mathrm{r}$ and $\mathbf{k}_\mathrm{t}$ are perpendicular.

The angle that satisfies this equation, in terms of the refractive indices, is readily found to be

$$\theta_{\mathrm{B}} = \tan^{-1}\frac{n_{\mathrm{t}}}{n_{\mathrm{i}}} \tag{3.36}$$

We have replaced the specific θ_{i} with θ_{B} in honor of Sir David Brewster who first discovered the phenomenon. The angle θ_{B} is called *Brewster's angle*. At Brewster's angle, no p-polarized light reflects (see L 3.4). Physically, the p-polarized light cannot reflect because $\mathbf{k}_{\mathrm{r}}$ and $\mathbf{k}_{\mathrm{t}}$ are perpendicular. A reflection would require the microscopic dipoles at the surface of the second material to radiate along their axes, which they cannot do. Maxwell's equations 'know' about this, and so everything is nicely consistent.

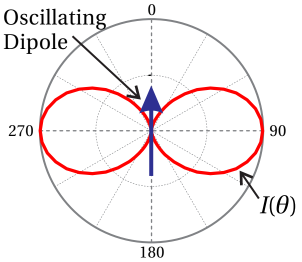


Figure 3.7 The intensity radiation pattern of an oscillating dipole as a function of angle. Note that the dipole does not radiate along the axis of oscillation, giving rise to Brewster's angle for reflection.

3.5 Total Internal Reflection

From Snell's law (3.7), we can compute the transmitted angle in terms of the incident angle:

$$\theta_{\mathrm{t}} = \sin^{-1}\left(\frac{n_{\mathrm{i}}}{n_{\mathrm{t}}}\sin\theta_{\mathrm{i}}\right) \tag{3.37}$$

The angle θ_{t} is real only if the argument of the inverse sine is less than or equal to one. If $n_{\mathrm{i}} > n_{\mathrm{t}}$, we can find a *critical angle* beyond which the argument begins to exceed one:

$$\theta_{\mathrm{c}} \equiv \sin^{-1}\frac{n_{\mathrm{t}}}{n_{\mathrm{i}}} \tag{3.38}$$

When $\theta_{\mathrm{i}} > \theta_{\mathrm{c}}$, then there is *total internal reflection* and we can directly show that $R_s = 1$ and $R_p = 1$ (see P3.9).[5] To demonstrate this, one computes the Fresnel coefficients (3.20) and (3.22) while employing the following substitution:

$$\cos\theta_{\mathrm{t}} = \sqrt{1-\sin^2\theta_{\mathrm{t}}} = i\sqrt{\frac{n_{\mathrm{i}}^2}{n_{\mathrm{t}}^2}\sin^2\theta_{\mathrm{i}} - 1} \qquad (\theta_{\mathrm{i}} > \theta_{\mathrm{c}}) \tag{3.39}$$

(see P0.19).

In this case, θ_{t} is a complex number. However, we do not assign geometrical significance to it in terms of any direction. Actually, we don't even need to know the value for θ_{t}; we need only the values for $\sin\theta_{\mathrm{t}}$ and $\cos\theta_{\mathrm{t}}$, as specified by Snell's law (3.7) and (3.39). Even though $\sin\theta_{\mathrm{t}}$ is greater than one and $\cos\theta_{\mathrm{t}}$ is imaginary, we can use their values to compute r_s, r_p, t_s, and t_p. (Complex notation is wonderful!)

Upon substitution of (3.39) into the Fresnel reflection coefficients (3.20) and (3.22) we obtain

$$r_s = \frac{n_{\mathrm{i}}\cos\theta_{\mathrm{i}} - i n_{\mathrm{t}}\sqrt{\frac{n_{\mathrm{i}}^2}{n_{\mathrm{t}}^2}\sin^2\theta_{\mathrm{i}} - 1}}{n_{\mathrm{i}}\cos\theta_{\mathrm{i}} + i n_{\mathrm{t}}\sqrt{\frac{n_{\mathrm{i}}^2}{n_{\mathrm{t}}^2}\sin^2\theta_{\mathrm{i}} - 1}} \qquad (\theta_{\mathrm{i}} > \theta_{\mathrm{c}}) \tag{3.40}$$

---

[5] M. Born and E. Wolf, *Principles of Optics*, 7th ed., Sect. 1.5.4 (Cambridge University Press, 1999).

and

$$r_p = -\frac{n_\mathrm{t}\cos\theta_\mathrm{i} - i\,n_\mathrm{i}\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i}-1}}{n_\mathrm{t}\cos\theta_\mathrm{i} + i\,n_\mathrm{i}\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i}-1}} \qquad (\theta_\mathrm{i} > \theta_\mathrm{c}) \quad (3.41)$$

These Fresnel coefficients can be manipulated (see P3.9) into the forms

$$r_s = \exp\left\{-2i\tan^{-1}\left[\frac{n_\mathrm{t}}{n_\mathrm{i}\cos\theta_\mathrm{i}}\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i}-1}\right]\right\} \qquad (\theta_\mathrm{i} > \theta_\mathrm{c}) \quad (3.42)$$

and

$$r_p = -\exp\left\{-2i\tan^{-1}\left[\frac{n_\mathrm{i}}{n_\mathrm{t}\cos\theta_\mathrm{i}}\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i}-1}\right]\right\} \qquad (\theta_\mathrm{i} > \theta_\mathrm{c}) \quad (3.43)$$

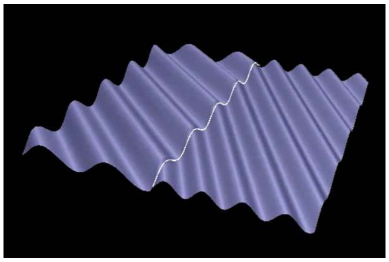


Figure 3.8 Animation of light waves incident on an interface both below and beyond the critical angle.

Each coefficient has a different phase (note $n_\mathrm{i}/n_\mathrm{t}$ vs. $n_\mathrm{t}/n_\mathrm{i}$ in the expressions), which means that the s- and p-polarized fields experience different phase shifts upon reflection. Nevertheless, we definitely have $|r_s| = 1$ and $\left|r_p\right| = 1$. We rightly conclude that 100% of the light reflects. The transmittance is zero as dictated by (3.30). We emphasize that one should not employ (3.32) or (3.33) in the case of total internal reflection, as the imaginary θ_t makes the geometric factor in this equation invalid.

Even with zero transmittance, the boundary conditions from Maxwell's equations (as worked out in appendix 3.A) require that the fields be nonzero on the transmitted side of the boundary, meaning $t_s \neq 0$ and $t_\mathrm{p} \neq 0$. While this situation may seem like a contradiction at first, it is an accurate description of what actually happens. The coefficients t_s and t_p characterize *evanescent waves* that exist on the transmitted side of the interface. The evanescent wave travels *parallel* to the interface so that no energy is conveyed away from the interface deeper into the medium on the transmission side.

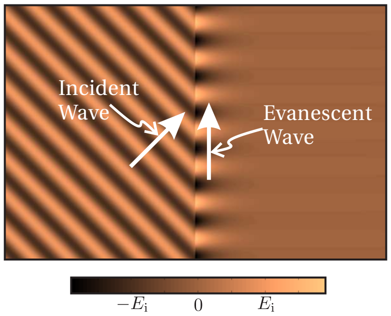


Figure 3.9 A wave experiencing total internal reflection creates an evanescent wave that propagates parallel to the interface. (The reflected wave is not shown.)

To compute the explicit form of the evanescent wave,[6] we plug (3.39) as well as Snell's law into the transmitted field (3.2):

$$\begin{aligned}\mathbf{E}_\mathrm{t} &= \left[E_\mathrm{t}^{(p)}\left(\hat{\mathbf{y}}\cos\theta_\mathrm{t} - \hat{\mathbf{z}}\sin\theta_\mathrm{t}\right) + \hat{\mathbf{x}}E_\mathrm{t}^{(s)}\right] e^{i[k_\mathrm{t}(y\sin\theta_\mathrm{t} + z\cos\theta_\mathrm{t}) - \omega t]} \\ &= \left[t_p E_\mathrm{i}^{(p)}\left(\hat{\mathbf{y}}i\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i}-1} - \hat{\mathbf{z}}\frac{n_\mathrm{i}}{n_\mathrm{t}}\sin\theta_\mathrm{i}\right) + \hat{\mathbf{x}}t_s E_\mathrm{i}^{(s)}\right] e^{-k_\mathrm{t}z\sqrt{\frac{n_\mathrm{i}^2}{n_\mathrm{t}^2}\sin^2\theta_\mathrm{i}-1}}\, e^{i\left[k_\mathrm{t}y\frac{n_\mathrm{i}}{n_\mathrm{t}}\sin\theta_\mathrm{i} - \omega t\right]}\end{aligned} \quad (3.44)$$

Figure 3.9 plots the evanescent wave described by (3.44) along with the associated incident wave. The phase of the evanescent wave indicates that it propagates parallel to the boundary (in the y-dimension). Its strength decays exponentially away from the boundary (in the z-dimension). We leave the calculation of t_s and t_p as an exercise (P3.10).

---

[6] G. R. Fowles, *Introduction to Modern Optics*, 2nd ed., Sect 2.9 (New York: Dover, 1975).

3.6 Reflections from Metal

In this section we generalize our analysis to materials with complex refractive index $\mathcal{N} \equiv n + i\kappa$. As a reminder, the imaginary part of the index controls attenuation of a wave as it propagates within a material. The real part of the index governs the oscillatory nature of the wave. It turns out that both the imaginary and real parts of the index strongly influence the reflection of light from a surface. The reader may be grateful that there is no need to re-derive the Fresnel coefficients (3.20)–(3.23) for the case of complex indices. The coefficients remain valid whether the index is real or complex – just replace the real index n with the complex index $\mathcal{N}$. However, we do need to be a bit careful when applying them.

We restrict our discussion to *reflections* from a metallic or other absorbing material surface. As we found in the case of total internal reflection, we actually do not need to know the transmitted angle θ_t to employ Fresnel reflection coefficients (3.20) and (3.22). We need only acquire expressions for $\cos\theta_t$ and $\sin\theta_t$, and we can obtain those from Snell's law (3.7). To minimize complications, we let the incident refractive index be $n_i = 1$ (which is often the case). Let the index on the transmitted side be written as $\mathcal{N}_t = \mathcal{N}$. Then by Snell's law, the sine of the transmitted angle is

$$\sin\theta_t = \frac{\sin\theta_i}{\mathcal{N}} \tag{3.45}$$

This expression is of course complex since $\mathcal{N}$ is complex, which is just fine.[7] The cosine of the same angle is

$$\cos\theta_t = \sqrt{1-\sin^2\theta_t} = \frac{1}{\mathcal{N}}\sqrt{\mathcal{N}^2-\sin^2\theta_i} \tag{3.46}$$

The positive sign in front of the square root is appropriate since it is clearly the right choice if the imaginary part of the index approaches zero.

Upon substitution of these expressions, the Fresnel reflection coefficients (3.20) and (3.22) become

$$r_s = \frac{\cos\theta_i - \sqrt{\mathcal{N}^2-\sin^2\theta_i}}{\cos\theta_i + \sqrt{\mathcal{N}^2-\sin^2\theta_i}} \tag{3.47}$$

and

$$r_p = \frac{\sqrt{\mathcal{N}^2-\sin^2\theta_i} - \mathcal{N}^2\cos\theta_i}{\sqrt{\mathcal{N}^2-\sin^2\theta_i} + \mathcal{N}^2\cos\theta_i} \tag{3.48}$$

These expressions are tedious to evaluate. When evaluating the expressions, it is usually desirable to put them into the form

$$r_s = |r_s|\, e^{i\phi_s} \tag{3.49}$$

and

$$r_p = \left|r_p\right| e^{i\phi_p} \tag{3.50}$$

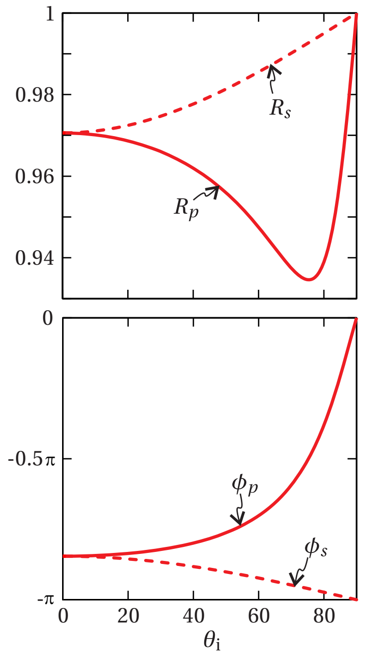


Figure 3.10 The reflectances (top) with associated phases (bottom) for silver, which has index $n = 0.13$ and $\kappa = 4.0$ near $\lambda = 633$ nm. Note the minimum of R_p corresponding to a kind of Brewster's angle.

---

[7] See M. Born and E. Wolf, *Principles of Optics*, 7th ed., Sect. 14.2 (Cambridge University Press, 1999).

We refrain from putting (3.47) and (3.48) into this form using the general expressions; we would get a big mess. It is a good idea to let your calculator or a computer do it after a specific value for $\mathcal{N} \equiv n + i\kappa$ is chosen. An important point to notice is that the phases upon reflection can be very different for s and p-polarization components (i.e. ϕ_p and ϕ_s can be very different). This is true in general, even when the reflectivity is high (i.e. $|r_s|$ and $\left|r_\mathrm{p}\right|$ on the order of unity).

Brewster's angle exists also for surfaces with complex refractive index. However, in general the expressions (3.48) and (3.50) do not go to zero at any incident angle θ_i. Rather, the reflection of p-polarized light can go through a minimum at some angle θ_i, which we refer to as Brewster's angle (see Fig. 3.10). This minimum is best found numerically since the general expression for $\left|r_p\right|$ in terms of n and κ and as a function of θ_i can be unwieldy.

Appendix 3.A Boundary Conditions For Fields at an Interface

We are interested in the continuity of fields across a boundary from one medium with index n_1 to another medium with index n_2. We will show that the components of electric field and the magnetic field parallel to the interface surface must be the same on either side (adjacent to the interface). This result is independent of the refractive index of the materials; in the case of the magnetic field we assume the permeability μ_0 is the same on both sides. To derive the boundary conditions, we consider a surface S (a rectangle) that is *perpendicular* to the interface between the two media and which extends into both media, as depicted in Fig. 3.11.

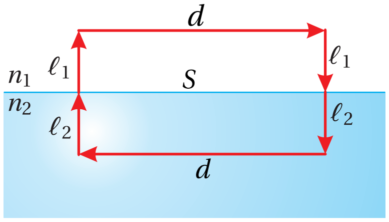


Figure 3.11 Interface of two materials.

First we examine the integral form of Faraday's law (1.14)

$$\oint_C \mathbf{E}\cdot d\boldsymbol{\ell} = -\frac{\partial}{\partial t}\int_S \mathbf{B}\cdot\hat{\mathbf{n}}\, da \tag{3.51}$$

applied to the rectangular contour depicted in Fig. 3.11. We perform the path integration on the left-hand side around the loop as follows:

$$\oint \mathbf{E}\cdot d\boldsymbol{\ell} = E_{1||}d - E_{1\perp}\ell_1 - E_{2\perp}\ell_2 - E_{2||}d + E_{2\perp}\ell_2 + E_{1\perp}\ell_1 = \left(E_{1||} - E_{2||}\right)d \tag{3.52}$$

Here, $E_{1||}$ refers to the component of the electric field in the material with index n_1 that is parallel to the interface. $E_{1\perp}$ refers to the component of the electric field in the material with index n_1 which is perpendicular to the interface. Similarly, $E_{2||}$ and $E_{2\perp}$ are the parallel and perpendicular components of the electric field in the material with index n_2. We have assumed that the rectangle is small enough that the fields are uniform within the half rectangle on either side of the boundary.

Next, we shrink the loop down until it has zero surface area by letting the lengths ℓ_1 and ℓ_2 go to zero. In this situation, the right-hand side of Faraday's law (3.51) goes to zero

$$\int_S \mathbf{B}\cdot\hat{\mathbf{n}}\, da \to 0 \tag{3.53}$$

and we are left with

$$E_{1||} = E_{2||} \tag{3.54}$$

This simple relation is a general boundary condition, which is met at any material interface. The component of the electric field that lies in the plane of the interface must be the same on both sides of the interface.

We now derive a similar boundary condition for the magnetic field using the integral form of Ampere's law:[8]

$$\oint_C \mathbf{B} \cdot d\ell = \mu_0 \int_S \left(\mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right) \cdot \hat{\mathbf{n}}\, da \tag{3.55}$$

As before, we are able to perform the path integration on the left-hand side for the geometry depicted in the figure, which gives

$$\oint \mathbf{B} \cdot d\ell = B_{1||} d - B_{1\perp} \ell_1 - B_{2\perp} \ell_2 - B_{2||} d + B_{2\perp} \ell_2 + B_{1\perp} \ell_1 = \left(B_{1||} - B_{2||} \right) d \tag{3.56}$$

The notation for parallel and perpendicular components on either side of the interface is similar to that used in (3.52).

Again, we can shrink the loop down until it has zero surface area by letting the lengths ℓ_1 and ℓ_2 go to zero. In this situation, the right-hand side of (3.55) goes to zero (ignoring the possibility of surface currents):

$$\int_S \left(\mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right) \cdot \hat{\mathbf{n}}\, da \to 0 \tag{3.57}$$

and we are left with

$$B_{1||} = B_{2||} \tag{3.58}$$

This is a general boundary condition that must be satisfied at the material interface.

---

[8] This form can be obtained from (1.4) by integration over the surface S in Fig. 3.11 and applying Stokes' theorem (0.12) to the magnetic field term.

Exercises

Exercises for 3.2 The Fresnel Coefficients

P3.1 Derive the Fresnel coefficients (3.22) and (3.23) for p-polarized light.

P3.2 Verify that each of the alternative forms given in (3.20)–(3.23) are equivalent. Show that at normal incidence (i.e. $\theta_\mathrm{i} = \theta_\mathrm{t} = 0$) the Fresnel coefficients reduce to

$$\lim_{\theta_\mathrm{i}\to 0} r_s = \lim_{\theta_\mathrm{i}\to 0} r_p = -\frac{n_\mathrm{t} - n_\mathrm{i}}{n_\mathrm{t} + n_\mathrm{i}} \quad \text{and} \quad \lim_{\theta_\mathrm{i}\to 0} t_s = \lim_{\theta_\mathrm{i}\to 0} t_p = \frac{2n_\mathrm{i}}{n_\mathrm{t} + n_\mathrm{i}}$$

HINT: Substitute from Snell's law.

P3.3 Use a computer to make a plot similar to Fig. 3.3 of r_p, t_p, r_s, t_s as a function of the incident angle for an air-diamond interface. Use $n_\mathrm{i} = 1$ for air and $n_\mathrm{t} = 2.42$ for diamond. Note Brewster's angle where r_p goes through zero.

Exercises for 3.3 Reflectance and Transmittance

L3.4 (a) In the laboratory, measure the reflectance for both s and p polarized light from a flat glass surface at about ten angles. Especially watch for Brewster's angle (described in section 3.4). You can normalize the detector by measuring the beam before the glass surface. Figure 3.12 illustrates the experimental setup. (video)

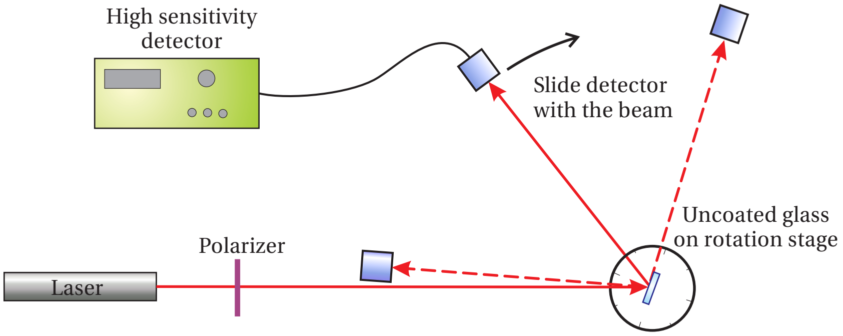

High sensitivity detector
Slide detector with the beam
Uncoated glass on rotation stage
Polarizer
Laser

Figure 3.12 Experimental setup for lab 3.4.

(b) Use a computer to calculate the theoretical air-to-glass reflectance as a function of incident angle (i.e. plot R_s and R_p as a function of θ_i). Take the index of refraction for glass to be $n_\mathrm{t} = 1.54$ and the index for air to be one. Plot this theoretical calculation as a smooth line on a graph. Plot your experimental data from (a) as points on this same graph (not points connected by lines).

P3.5 A pentaprism is a five-sided reflecting prism used to deviate a beam of light by 90° without inverting an image (see Fig. 3.13). Pentaprisms are used in the viewfinders of SLR cameras.

(a) What prism angle β is required for a normal-incidence beam from the left to exit the bottom surface at normal incidence?

(b) If all interfaces of the pentaprism are uncoated glass with index $n = 1.5$, what fraction of the intensity would get through this system for a normal incidence beam? Compute for p-polarized light, and include transmission through the first and final surfaces as well as reflection at the two interior surfaces.

NOTE: You will find that the overall transmission through the device is very poor. The reflecting surfaces on pentaprisms are usually treated with a high-reflection coating and the transmitting surfaces are treated with anti-reflection coatings.

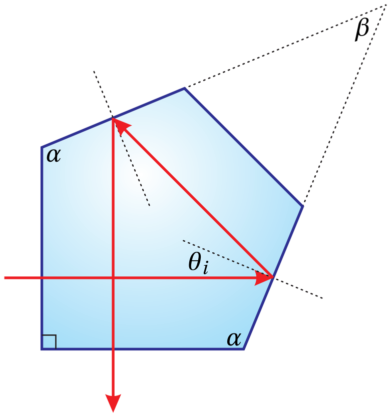


Figure 3.13

P3.6 (a) Show *analytically* for s-polarized light that $R_s + T_s = 1$, where R_s is given by (3.26) and T_s is given by (3.33).

(b) Repeat for p-polarized light.

Exercises for 3.4 Brewster's Angle

P3.7 (a) Find Brewster's angle for an air-glass interface with $n_{\text{glass}} = 1.5$.

(b) Compute R_s and R_p at this angle.

Exercises for 3.5 Total Internal Reflection

P3.8 Diamonds have an index of refraction of $n = 2.42$ which allows total internal reflection to occur at relatively shallow angles of incidence. Gem cutters choose facet angles that ensure most of the light entering the top of the diamond will reflect back out to give the stone its expensive sparkle. One such cut, the "Eulitz Brilliant" cut, is shown in Fig. 3.14.

(a) What is the critical angle for diamond?

(b) What fraction of the light reflects for internal angles $\theta_i = 40.5°$ and $\theta_i = 50.6°$? One way to spot a fake diamond is by noticing reduced brilliance in the sparkle. Are these angles both beyond the critical angle for fused quartz ($n = 1.46$)?

(c) For each angle and assuming s-polarized light, find the phase shift upon reflection ϕ_s where $r_s = |r_s|\, e^{i\phi_s}$.

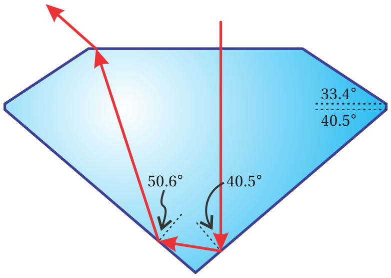


Figure 3.14 A Eulitz Brilliant cut diamond.

P3.9 Derive (3.42) and (3.43) and show that $R_s = 1$ and $R_p = 1$. HINT: See problem P0.15.

P3.10 Develop expressions for t_s and t_p in terms of θ_i and the refractive indices in the case of total internal reflection. Put your answer in polar form (i.e. $t = |t|e^{i\phi}$).

P3.11 Use a computer to plot the air-to-water transmittance (both T_s and T_p) as a function of incident angle (i.e. plot (3.30) as a function of θ_i). On a separate graph, plot the water-to-air transmittance. Take the index of air to be one; the index of refraction for water is $n = 1.33$.

P3.12 Light ($\lambda_\mathrm{vac} = 500$ nm) reflects internally from a glass surface ($n = 1.5$) surrounded by air. The incident angle is $\theta_\mathrm{i} = 45°$. An evanescent wave travels parallel to the surface on the air side. At what distance from the surface is the amplitude of the evanescent wave $1/e$ of its value at the surface?

Exercises for 3.6 Reflections from Metal

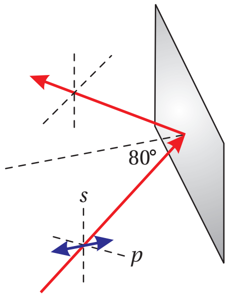


Figure 3.15 Geometry for P3.13

P3.13 The complex index for silver is given by $n = 0.13$ and $\kappa = 4.0$.[9] Find r_s and r_p when reflecting at $\theta_\mathrm{i} = 80°$ and put them into the forms (3.49) and (3.50). Assume the light propagates in vacuum on the incident side.

Answer: $r_s = 0.997e^{-i3.057}$, $r_p = 0.969e^{-i1.187}$

P3.14 (a) Using a computer, plot R_s, R_p versus θ_i for silver ($n = 0.13$ and $\kappa = 4.0$). Make a separate plot of the phases ϕ_s and ϕ_p from (3.49) and (3.50). Clearly label each plot.

(b) Can you identify Brewster's angle (i.e. where R_p is minimum)?

---

[9] *Handbook of Optical Constants of Solids*, Edited by E. D. Palik (Elsevier, 1997).

Chapter 4

Multiple Parallel Interfaces

In chapter 3, we studied the transmission and reflection of light at a single interface between two (isotropic homogeneous) materials with indices n_i and n_t. We found that the percent of light reflected versus transmitted depends on the incident angle and on whether the light is s- or p-polarized. The Fresnel coefficients r_s, t_s, r_p, t_p (3.20)–(3.23) connect the reflected and transmitted *fields* to the incident field. Similarly, either R_s and T_s or R_p and T_p determine the fraction of incident *power* that either reflects or transmits (see (3.26) and (3.30)).

In this chapter we consider the overall transmission and reflection through multiple parallel interfaces. We start with a two-interface system, where a layer of material is inserted between the initial and final materials. This situation occurs frequently in optics. For example, lenses are often coated with a thin layer of material in an effort to reduce reflections. Metal mirrors usually have a thin oxide layer or a protective coating between the metal and the air. We can develop reflection and transmission coefficients r^tot and t^tot, which apply to the overall double-boundary system, similar to the Fresnel coefficients for a single boundary. Likewise, we can compute an overall reflectance and transmittance R^tot and T^tot. These can be used to compute the 'tunneling' of evanescent waves across a gap between two parallel surfaces when the critical angle for total internal reflection is exceeded.

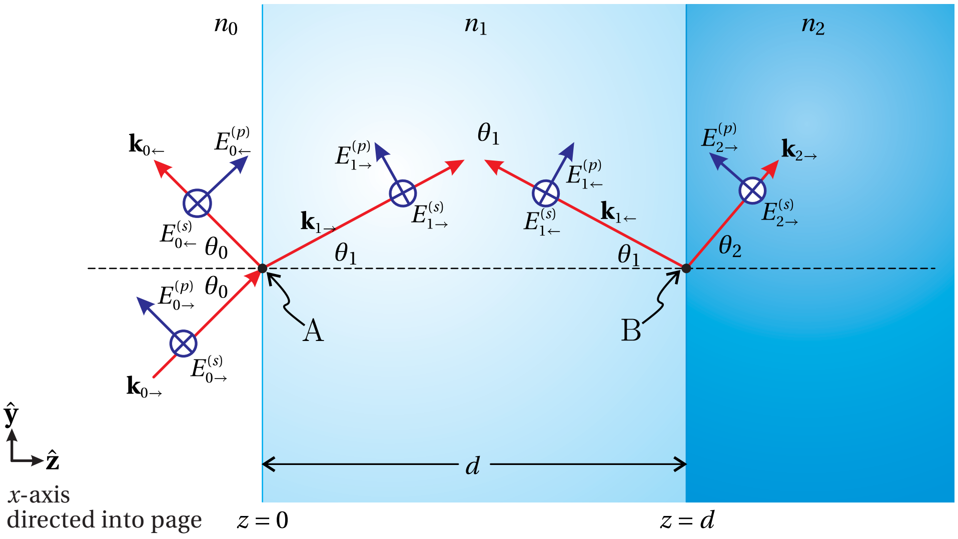
The formalism we develop for the double-boundary problem is useful for describing a simple instrument called a *Fabry-Perot etalon* (or *interferometer* if the instrument has the capability of variable spacing between the two surfaces). Such an instrument, which is constructed from two partially reflective parallel surfaces, is useful for distinguishing closely spaced wavelengths.

Finally, in this chapter we will extend our analysis to *multilayer coatings*, where an arbitrary number of interfaces exist between many material layers. Multilayers are often used to make highly reflective mirror coatings from dielectric materials (as opposed to metallic materials). Such mirror coatings can reflect with efficiencies greater than 99.9% at specified wavelengths. In contrast, metallic mirrors typically reflect with ~ 96% efficiency, which can be a significant loss if there are many mirrors in an optical system. Dielectric multilayer coatings also have the advantage of being more durable and less prone to damage from high-intensity lasers.

4.1 Double-Interface Problem Solved Using Fresnel Coefficients

Consider a slab of material sandwiched between two other materials as depicted in Fig. 4.1. Because there are multiple reflections inside the middle layer, we have dropped the subscripts i, r, and t used in chapter 3 and instead use the symbols $\rightarrow$ and $\leftarrow$ to indicate forward and backward traveling waves, respectively. Let n_1 stand for the refractive index of the middle layer. For consistency with notation that we will later use for many-layer systems, let n_0 and n_2 represent the indices of the other two regions. For simplicity, we assume that indices are real. As with the single-boundary problem, we are interested in finding the overall transmitted fields $E_{2\rightarrow}^{(s)}$ and $E_{2\rightarrow}^{(p)}$ and the overall reflected fields $E_{0\leftarrow}^{(s)}$ and $E_{0\leftarrow}^{(p)}$ in terms of the incident fields $E_{0\rightarrow}^{(s)}$ and $E_{0\rightarrow}^{(p)}$.

Both forward and backward traveling plane waves exist in the middle region. Our intuition rightly tells us that in this region there are many reflections, bouncing both forward and backward between the two surfaces. It might therefore seem that we need to keep track of an infinite number of plane waves, each corresponding to a different number of bounces. Fortunately, the many forward-traveling plane waves all travel in the same direction. Similarly, the backward-traveling plane waves are all parallel. These plane-wave fields then join neatly into a single net forward-moving and a single net backward-moving plane wave within the



n_0
n_1
n_2
$\mathbf{k}_{0\leftarrow}$
$E_{0\leftarrow}^{(p)}$
$E_{1\rightarrow}^{(p)}$
θ_1
$E_{1\leftarrow}^{(p)}$
$E_{2\rightarrow}^{(p)}$
$\mathbf{k}_{2\rightarrow}$
$\mathbf{k}_{1\rightarrow}$
$E_{1\rightarrow}^{(s)}$
$E_{1\leftarrow}^{(s)}$
$\mathbf{k}_{1\leftarrow}$
$E_{2\rightarrow}^{(s)}$
$E_{0\leftarrow}^{(s)}$
θ_0
θ_1
θ_1
θ_2
$E_{0\rightarrow}^{(p)}$
θ_0
A
B
$E_{0\rightarrow}^{(s)}$
$\mathbf{k}_{0\rightarrow}$
$\hat{\mathbf{y}}$
$\hat{\mathbf{z}}$
x-axis directed into page
d
$z = 0$
$z = d$

Figure 4.1 Waves propagating through a dual interface between materials.

middle region.[1]

As of yet, we do not know the amplitudes or phases of the net forward and net backward traveling plane waves in the middle layer. We denote them by $E_{1\rightarrow}^{(s)}$ and $E_{1\leftarrow}^{(s)}$ or by $E_{1\rightarrow}^{(p)}$ and $E_{1\leftarrow}^{(p)}$, separated into their s and p components as usual. Similarly, $E_{0\leftarrow}^{(s)}$ and $E_{0\leftarrow}^{(p)}$ as well as $E_{2\rightarrow}^{(s)}$ and $E_{2\rightarrow}^{(p)}$ are understood to include light that 'leaks' through the boundaries from the middle region. Thus, we need only concern ourselves with the five plane waves depicted in Fig. 4.1.

The various plane-wave fields are connected to each other at the boundaries via the single-boundary Fresnel coefficients (3.20)–(3.23). At the first surface we define

$$\begin{aligned} r_s^{0\to1} &\equiv \frac{\sin\theta_1\cos\theta_0 - \sin\theta_0\cos\theta_1}{\sin\theta_1\cos\theta_0 + \sin\theta_0\cos\theta_1} & \qquad r_p^{0\to1} &\equiv \frac{\sin\theta_1\cos\theta_1 - \sin\theta_0\cos\theta_0}{\sin\theta_1\cos\theta_1 + \sin\theta_0\cos\theta_0} \\ t_s^{0\to1} &\equiv \frac{2\sin\theta_1\cos\theta_0}{\sin\theta_1\cos\theta_0 + \sin\theta_0\cos\theta_1} & \qquad t_p^{0\to1} &\equiv \frac{2\sin\theta_1\cos\theta_0}{\sin\theta_1\cos\theta_1 + \sin\theta_0\cos\theta_0} \end{aligned} \tag{4.1}$$

The notation 0 → 1 indicates the first surface from the perspective of starting on the incident side and propagating towards the middle layer. The Fresnel coefficients for the backward traveling light approaching the first interface from *within* the middle layer are given by

$$\begin{aligned} r_s^{0\leftarrow1} &= -r_s^{0\to1} & \qquad r_p^{0\leftarrow1} &= -r_p^{0\to1} \\ t_s^{0\leftarrow1} &\equiv \frac{2\sin\theta_0\cos\theta_1}{\sin\theta_0\cos\theta_1 + \sin\theta_1\cos\theta_0} & \qquad t_p^{0\leftarrow1} &\equiv \frac{2\sin\theta_0\cos\theta_1}{\sin\theta_0\cos\theta_0 + \sin\theta_1\cos\theta_1} \end{aligned} \tag{4.2}$$

where 0 ← 1 indicates connections at the first interface, but from the perspective of beginning inside the middle layer. Finally, the single-boundary coefficients for light approaching the second interface are

$$\begin{aligned} r_s^{1\to2} &\equiv \frac{\sin\theta_2\cos\theta_1 - \sin\theta_1\cos\theta_2}{\sin\theta_2\cos\theta_1 + \sin\theta_1\cos\theta_2} & \qquad r_p^{1\to2} &\equiv \frac{\sin\theta_2\cos\theta_2 - \sin\theta_1\cos\theta_1}{\sin\theta_2\cos\theta_2 + \sin\theta_1\cos\theta_1} \\ t_s^{1\to2} &\equiv \frac{2\sin\theta_2\cos\theta_1}{\sin\theta_2\cos\theta_1 + \sin\theta_1\cos\theta_2} & \qquad t_p^{1\to2} &\equiv \frac{2\sin\theta_2\cos\theta_1}{\sin\theta_2\cos\theta_2 + \sin\theta_1\cos\theta_1} \end{aligned} \tag{4.3}$$

In a similar fashion, the notation 1 → 2 indicates connections made at the second interface from the perspective of beginning in the middle layer.

To solve for the connections between the five fields depicted in Fig.4.1, we will need four equations for either s or p polarization (taking the incident field as a given). To simplify things, we will consider s-polarized light in the upcoming analysis. The equations for p-polarized light look exactly the same; just replace the subscript s with p. Through the remainder of this section and the next, we will continue to economize by writing the equations only for s-polarized light with the understanding that they apply equally well to p-polarized light.

The forward-traveling wave in the middle region arises from both a transmission of the incident wave and a reflection of the backward-traveling wave in the middle region at the first interface. Using the Fresnel coefficients, we can write $E_{1\rightarrow}^{(s)}$ as the sum of fields arising from $E_{0\rightarrow}^{(s)}$ and $E_{1\leftarrow}^{(s)}$ as follows:

---

[1] The sum of parallel plane waves $\sum_j \mathbf{E}_j e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$, where the phase of each wave is contained in $\mathbf{E}_j$, can be written as $(\sum_j \mathbf{E}_j)e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$, which is effectively one plane wave.

$$E_{1\rightarrow}^{(s)} = t_s^{0\rightarrow 1} E_{0\rightarrow}^{(s)} + r_s^{0\leftarrow 1} E_{1\leftarrow}^{(s)} \tag{4.4}$$

The factor $t_s^{0\rightarrow 1}$ and $r_s^{0\leftarrow 1}$ are the single-boundary Fresnel coefficients selected from (4.1). Similarly, the overall reflected field $E_{0\leftarrow}^{(s)}$, is given by the reflection of the incident field and the transmission of the backward-traveling field in the middle region according to

$$E_{0\leftarrow}^{(s)} = r_s^{0\rightarrow 1} E_{0\rightarrow}^{(s)} + t_s^{0\leftarrow 1} E_{1\leftarrow}^{(s)} \tag{4.5}$$

Two connections made; two to go.

Before we continue, we need to specify an origin so that we can calculate phase shifts associated with propagation in the middle region. Propagation was not an issue in the single-boundary problem studied back in chapter 3. However, in the double-boundary problem, the thickness of the middle region dictates phase variations that strongly influence the result. We take the origin to be located on the first interface, as shown in Fig. 4.1. Since all fields in (4.4) and (4.5) are evaluated at the origin $(y, z) = (0, 0)$, there were no phase factors needed.

We will connect the plane-wave fields across the second interface at the point $\mathbf{r} = \hat{\mathbf{z}}d$. The appropriate phase-adjusted[2] field at $(y, z) = (0, d)$ is $E_{1\rightarrow}^{(s)} e^{i\mathbf{k}_{1\rightarrow}\cdot\mathbf{r}} = E_{1\rightarrow}^{(s)} e^{ik_1 d\cos\theta_1}$, since $E_{1\rightarrow}^{(s)}$ is the field at the origin $(y, z) = (0, 0)$. The transmitted field in the final medium arises only from the forward-traveling field in the middle region, and at our selected point it is

$$E_{2\rightarrow}^{(s)} = t_s^{1\rightarrow 2} E_{1\rightarrow}^{(s)} e^{ik_1 d\cos\theta_1} \tag{4.6}$$

Note that $E_{2\rightarrow}^{(s)}$ stand for the transmitted field at the point $(y, z) = (0, d)$; its local phase can be built into its definition so no need to write an explicit phase.

The backward-traveling plane wave in the middle region arises from the reflection of the forward-traveling plane wave in that region:

$$E_{1\leftarrow}^{(s)} = E_{1\rightarrow}^{(s)} e^{ik_1 d\cos\theta_1} r_s^{1\rightarrow 2} e^{ik_1 d\cos\theta_1} \tag{4.7}$$

We have written the phase terms on the right of (4.7) in a long form to emphasize that they describe a transmission through the middle layer, followed by a reflection from the second interface, and then another transmission through the middle layer back to the first interface.

The relations (4.4)–(4.7) permit us to find overall transmission and reflection coefficients for the two-interface problem.

Example 4.1

Derive the transmission coefficient that connects the final transmitted field to the incident field for the double-interface problem according to $t_s^{\text{tot}} \equiv E_{2\rightarrow}^{(s)} / E_{0\rightarrow}^{(s)}$.

---

[2] In the middle region, $\mathbf{k}_{1\rightarrow} = k_1 \left(\hat{\mathbf{y}}\sin\theta_1 + \hat{\mathbf{z}}\cos\theta_1\right)$ and $\mathbf{k}_{1\leftarrow} = k_1 \left(\hat{\mathbf{y}}\sin\theta_1 - \hat{\mathbf{z}}\cos\theta_1\right)$.

Solution: From (4.6) we may write

$$E_{1\rightarrow}^{(s)} = \frac{E_{2\rightarrow}^{(s)}}{t_s^{1\rightarrow 2}} e^{-ik_1 d\cos\theta_1} \tag{4.8}$$

Substitution of this into (4.7) gives

$$E_{1\leftarrow}^{(s)} = E_{2\rightarrow}^{(s)} \frac{r_s^{1\rightarrow 2}}{t_s^{1\rightarrow 2}} e^{ik_1 d\cos\theta_1} \tag{4.9}$$

Next, substituting both (4.8) and (4.9) into (4.4) yields the connection we seek between the incident and transmitted fields:

$$\frac{E_{2\rightarrow}^{(s)}}{t_s^{1\rightarrow 2}} e^{-ik_1 d\cos\theta_1} = t_s^{0\rightarrow 1} E_{0\rightarrow}^{(s)} + r_s^{0\leftarrow 1} E_{2\rightarrow}^{(s)} \frac{r_s^{1\rightarrow 2}}{t_s^{1\rightarrow 2}} e^{ik_1 d\cos\theta_1} \tag{4.10}$$

After rearranging, we arrive at the more useful form

$$t_s^{\text{tot}} \equiv \frac{E_{2\rightarrow}^{(s)}}{E_{0\rightarrow}^{(s)}} = \frac{t_s^{0\rightarrow 1} e^{ik_1 d\cos\theta_1} t_s^{1\rightarrow 2}}{1 - r_s^{0\leftarrow 1} r_s^{1\rightarrow 2} e^{2ik_1 d\cos\theta_1}} \tag{4.11}$$

(p can be switched for s)

The coefficient t_s^{tot} derived in Example 4.1 connects the amplitude and phase of the incident field to the amplitude and phase of the transmitted field in a manner similar to the single-boundary Fresnel coefficients. The numerator of (4.11) reminds us of the physics of the situation: the field transmits through the first interface, acquires a phase due to propagating through the middle layer, and then transmits through the second interface. The denominator of (4.11) modifies the result to account for feedback from multiple reflections in the middle region.[3]

The overall reflection coefficient is found to be (see P4.1)

$$r_s^{\text{tot}} \equiv \frac{E_{0\leftarrow}^{(s)}}{E_{0\rightarrow}^{(s)}} = r_s^{0\rightarrow 1} + \frac{t_s^{0\rightarrow 1} e^{ik_1 d\cos\theta_1} r_s^{1\rightarrow 2} e^{ik_1 d\cos\theta_1} t_s^{0\leftarrow 1}}{1 - r_s^{0\leftarrow 1} r_s^{1\rightarrow 2} e^{i2k_1 d\cos\theta_1}} \quad \text{(can switch } p \text{ for } s\text{)} \tag{4.12}$$

The initial reflection from the first interface is described by the first term $r_s^{0\rightarrow 1}$. The numerator in (4.12) can be simplified algebraically, but we have left it in this longer form to emphasize the physics of the situation: light transmits through the first interface, propagates through the middle layer, reflects from the second interface, propagates back through the middle layer, and transmits back through the first interface to interfere with the initial reflection. The denominator of the second term accounts for the effects of multiple-reflection feedback.

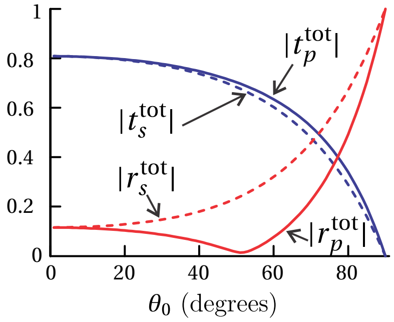


Figure 4.2 Plots of the magnitudes of the overall reflection and transmission coefficients for a quarter-wave thickness ($k_1 d = \pi/2$) of MgF$_2$ ($n1 = 1.38$) on glass ($n_2 = 1.5$) in air ($n_0 = 1$).

Figure 4.2 shows the magnitudes of the overall reflection and transmission coefficients for the case of a quarter-wave thickness coating of magnesium fluoride on glass with $k_1 d = \pi/2$. This coating is meant to reduce reflections by having the initial reflection described by the first term in (4.12) and the secondary reflection described by the second term add out of phase (i.e. have a relative phase shift of

---

[3] Our derivation method avoids the need for explicit accounting of multiple reflections. For an alternative approach arriving at the same result via an infinite geometric series, see M. Born and E. Wolf, *Principles of Optics*, 7th ed., Sect. 7.6.1 (Cambridge University Press, 1999) or G. R. Fowles, *Introduction to Modern Optics*, 2nd ed., Sect 4.1 (New York: Dover, 1975).

π). While this coating reduces the overall reflection as compared to an uncoated optic, note that it does not eliminate the reflection because the two interfering plane waves have different amplitudes. Figure 4.3 shows the phase of the overall reflection and transmission coefficients, written in the form $r_s^{\text{tot}} = |r_s^{\text{tot}}| e^{i\phi_{r_s}}$. At high incidence angles the s- and p-polarization reflection coefficients experience markedly different phase shifts.

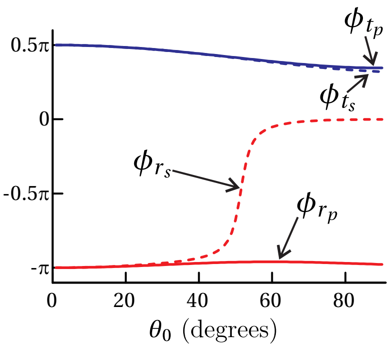


Figure 4.3 Plots of the phases of the overall reflection and transmission coefficients for a quarter-wave thickness ($k_1 d = \pi/2$) of MgF$_2$ ($n1 = 1.38$) on glass ($n_2 = 1.5$) in air ($n_0 = 1$).

4.2 Transmittance through Double Interface at Subcritical Angles

We are now in a position to calculate the fraction of power that transmits through or reflects from a double-interface arrangement. Because the transmission coefficient (4.11) has a simpler form than the reflection coefficient (4.12), it is easier to calculate the total transmittance T_s^{tot} and obtain the reflectance, if desired, from the relationship (see (3.30))

$$T_s^{\text{tot}} + R_s^{\text{tot}} = 1 \tag{4.13}$$

When the transmitted angle θ_2 is real (i.e. θ_1 does not exceed the critical angle), we may write the fraction of the transmitted power as in (3.33):

(p can be switched for s)

$$\begin{aligned} T_s^{\text{tot}} &= \frac{n_2 \cos\theta_2}{n_0 \cos\theta_0} \left| t_s^{\text{tot}} \right|^2 \\ &= \frac{n_2 \cos\theta_2}{n_0 \cos\theta_0} \frac{\left| t_s^{0\to1} \right|^2 \left| t_s^{1\to2} \right|^2}{\left| e^{-ik_1 d \cos\theta_1} - r_s^{0\leftarrow1} r_s^{1\to2} e^{ik_1 d \cos\theta_1} \right|^2} \end{aligned} \qquad (\theta_2 \text{ real}) \tag{4.14}$$

Note that we multiplied the numerator and denominator of (4.11) by $e^{-ik_1 d \cos\theta_1}$ before inserting it into (4.14), which make the denominator more symmetric for later convenience.

When θ_1 is also real (i.e. θ_0 also does not exceed the critical angle), we can simplify (4.14) into the following useful form (see P4.3):[4]

(p can be switched for s)

$$T_s^{\text{tot}} = \frac{T_s^{\max}}{1 + F_s \sin^2\left(\frac{\Phi_s}{2}\right)} \qquad (\theta_1 \text{ and } \theta_2 \text{ real}) \tag{4.15}$$

where

$$T_s^{\max} \equiv \frac{T_s^{0\to1} T_s^{1\to2}}{\left(1 - \sqrt{R_s^{0\leftarrow1} R_s^{1\to2}}\right)^2} \tag{4.16}$$

$$\Phi_s \equiv 2k_1 d \cos\theta_1 + \phi_{r_s^{0\leftarrow1}} + \phi_{r_s^{1\to2}} \tag{4.17}$$

and

$$F_s \equiv \frac{4\sqrt{R_s^{0\leftarrow1} R_s^{1\to2}}}{\left(1 - \sqrt{R_s^{0\leftarrow1} R_s^{1\to2}}\right)^2} \tag{4.18}$$

---

[4] M. Born and E. Wolf, *Principles of Optics*, 7th ed., Sect. 7.6.1 (Cambridge University Press, 1999).

The quantity $T_s^{\max}$ is the maximum possible transmittance of power through the two surfaces. The single-interface transmittances ($T_s^{0\to1}$ and $T_s^{1\to2}$) and reflectances ($R_s^{0\leftarrow1}$ and $R_s^{1\to2}$) are calculated from the single-interface Fresnel coefficients in the usual way as described in chapter 3. The numerator of $T_s^{\max}$ represents the combined transmittances for the two interfaces without considering feedback due to multiple reflections. The denominator enhances this value to account for reinforcing feedback in the middle layer.

The exact argument of the sine function, Φ_s, can strongly influence the transmission. The term $2k_1 d\cos\theta_1$ represents the phase delay acquired during round-trip propagation in the middle region. The terms $\phi_{r_s^{0\leftarrow1}}$ and $\phi_{r_s^{1\to2}}$ account for possible phase shifts upon reflection from each interface. They are defined indirectly by writing the single-boundary Fresnel reflection coefficients in polar format:

$$r_s^{0\leftarrow1} = \left|r_s^{0\leftarrow1}\right| e^{i\phi_{r_s^{0\leftarrow1}}} \qquad \text{and} \qquad r_s^{1\to2} = \left|r_s^{1\to2}\right| e^{i\phi_{r_s^{1\to2}}} \tag{4.19}$$

If the indices of refraction in all regions are real, $\phi_{r_s^{0\leftarrow1}}$ and $\phi_{r_s^{1\to2}}$ take on values of either zero or π (i.e. the coefficients are positive or negative real numbers). When the indices are complex, other phase values are possible.

F_s is called the *coefficient of finesse*, which determines how strongly the transmittance is influenced when Φ_s is varied (for example, through varying d or the wavelength λ_{vac}).

Example 4.2

Consider a 'beam splitter' designed for s-polarized light incident on a substrate of glass ($n = 1.5$) at 45° as shown in Fig. 4.4. A thin coating of zinc sulfide ($n = 2.32$) is applied to the front of the glass to cause about half of the light to reflect. A magnesium fluoride ($n = 1.38$) coating is applied to the back surface of the glass to minimize reflections at that surface.[5] Each coating constitutes a separate double-interface problem. The front coating is deferred to problem P4.5. In this example, find the highest transmittance possible through the antireflection film at the back of the 'beam splitter' and the smallest possible $\tilde{d}$ that accomplishes this for light with wavelength $\lambda_{\text{vac}} = 633$ nm.

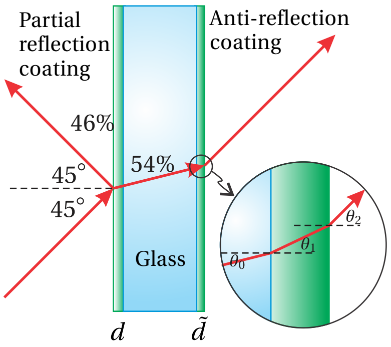


Figure 4.4 Side view of a beam-splitter.

Solution: For the back coating, we have $n_0 = 1.5$, $n_1 = 1.38$, and $n_2 = 1$. We can find θ_0 and θ_1 from $\theta_2 = 45°$ using Snell's law

$$n_1 \sin\theta_1 = \sin\theta_2 \qquad \Rightarrow \qquad \theta_1 = \sin^{-1}\left(\frac{\sin 45°}{1.38}\right) = 30.82°$$

$$n_0 \sin\theta_0 = \sin\theta_2 \qquad \Rightarrow \qquad \theta_0 = \sin^{-1}\left(\frac{\sin 45°}{1.5}\right) = 28.13°$$

---

[5] We ignore possible feedback between the front and rear coatings. Since the antireflection films are usually imperfect, beam splitter substrates are often slightly wedged so that unwanted reflections from the second surface travel in a different direction.

Next we calculate the single-boundary Fresnel coefficients:

$$r_s^{1\to 2} = -\frac{\sin(\theta_1 - \theta_2)}{\sin(\theta_1 + \theta_2)} = -\frac{\sin(30.82^\circ - 45^\circ)}{\sin(30.82^\circ + 45^\circ)} = 0.253$$

$$r_s^{0\leftarrow 1} = -\frac{\sin(\theta_1 - \theta_0)}{\sin(\theta_1 + \theta_0)} = -\frac{\sin(30.82^\circ - 28.13^\circ)}{\sin(30.82^\circ + 28.13^\circ)} = -0.0549$$

These coefficients give us the phase shift due to reflection

$$\phi_{r_s^{0\leftarrow 1}} = \pi\,, \quad \phi_{r_s^{1\to 2}} = 0$$

The single-boundary reflectances are given by

$$\begin{aligned} R_s^{0\leftarrow 1} &\equiv \left|r_s^{0\leftarrow 1}\right|^2 = |-0.0549|^2 = 0.0030 \\ R_s^{1\to 2} &\equiv \left|r_s^{1\to 2}\right|^2 = |0.253|^2 = 0.0640 \end{aligned}$$

and the transmittances are

$$\begin{aligned} T_s^{0\to 1} &= T_s^{0\leftarrow 1} = 1 - R_s^{0\leftarrow 1} = 1 - 0.0030 = 0.997 \\ T_s^{1\to 2} &= 1 - R_s^{1\to 2} = 1 - 0.0640 = 0.936 \end{aligned}$$

Finally, we calculate the coefficient of finesse

$$F = \frac{4\sqrt{R_s^{0\leftarrow 1} R_s^{1\to 2}}}{\left(1 - \sqrt{R_s^{0\leftarrow 1} R_s^{1\to 2}}\right)^2} = \frac{4\sqrt{(0.0030)\,(0.0640)}}{\left(1 - \sqrt{(0.0030)\,(0.0640)}\right)^2} = 0.0570$$

and the maximum transmittance

$$T_s^{\max} = \frac{T_s^{0\to 1} T_s^{1\to 2}}{\left(1 - \sqrt{R_s^{0\leftarrow 1} R_s^{1\to 2}}\right)^2} = \frac{(0.997)\,(0.936)}{\left(1 - \sqrt{(0.0030)\,(0.0640)}\right)^2} = 0.960$$

Putting everything together, we have

$$T_s^{\text{tot}} = \frac{0.960}{1 + 0.0570 \sin^2\left(\frac{2k_1 \tilde{d} \cos\theta_1 + \pi}{2}\right)}$$

The maximum transmittance occurs when the sine is zero. In that case, $T_s^{\text{tot}} = 0.960$, meaning that 96% of the light is transmitted. Without the coating, a situation we can recover by temporarily setting $\tilde{d} = 0$, the transmittance would be 90.8%, so the coating gives a significant improvement.

We find the smallest thickness $\tilde{d}$ that minimizes reflection by setting the argument of the sine to π:

$$2k_1 \tilde{d} \cos\theta_1 + \pi = 2\pi$$

Since $k_1 = 2\pi n_1 / \lambda_{\text{vac}}$, we have

$$\tilde{d} = \frac{\lambda_{\text{vac}}}{4n_1 \cos\theta_1} = \frac{633 \text{ nm}}{4(1.38)\cos 30.82^\circ} = 134 \text{ nm}$$

4.3 Beyond Critical Angle: Tunneling of Evanescent Waves

If $n_1 < n_0$, it is possible for θ_0 to exceed the critical angle at the first interface. In this case, (4.15) cannot be used to calculate transmittance. However, (4.14) still holds as long as the angle θ_2 is real (i.e. if the critical angle *in the absence of the middle layer* is not exceeded). In this case an evanescent wave occurs in the middle region, but not in the last region. If the second interface is sufficiently close to the first, the evanescent wave stimulates the second surface to produce a transmitted wave propagating at angle θ_2 in the last region. This behavior, called *tunneling* or *frustrated total internal reflection*, can be modeled using (4.14).

We do not need to deal directly with the complex angle θ_1. Rather, we just need $\sin\theta_1$ and $\cos\theta_1$ in order to calculate the single-boundary Fresnel coefficients. From Snell's law we have

$$\sin\theta_1 = \frac{n_0}{n_1}\sin\theta_0 = \frac{n_2}{n_1}\sin\theta_2 \tag{4.20}$$

even though $\sin\theta_1 > 1$. For the middle layer we write

$$\cos\theta_1 = i\sqrt{\sin^2\theta_1 - 1} \tag{4.21}$$

We illustrate how to apply (4.14) via a specific example:

Figure 4.5 Animation showing frustrated total internal reflection.

Example 4.3

Calculate the transmittance of p-polarized light through the region between two closely spaced 45° right prisms, as shown in Fig. 4.6, as a function of λ_{vac} and the prism spacing d. Take the index of refraction of the prisms to be $n = 1.5$ surrounded by index $n = 1$, and use $\theta_0 = \theta_2 = 45°$. Neglect possible reflections from the exterior surfaces of the prisms.

Solution: From (4.20) and (4.21) we have

$$\sin\theta_1 = 1.5\sin 45^\circ = 1.061 \qquad \text{and} \qquad \cos\theta_1 = i\sqrt{1.061^2 - 1} = i0.3536$$

We must compute various expressions involving Fresnel coefficients that appear in (4.14):

$$\left|t_p^{0\to1}\right|^2 = \left|\frac{2\cos\theta_0\sin\theta_1}{\cos\theta_1\sin\theta_1 + \cos\theta_0\sin\theta_0}\right|^2 = \left|\frac{2\frac{1}{\sqrt{2}}(1.061)}{(i0.3536)(1.061) + \frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}}}\right|^2 = 5.76$$

$$\left|t_p^{1\to2}\right|^2 = \left|\frac{2\cos\theta_1\sin\theta_2}{\cos\theta_2\sin\theta_2 + \cos\theta_1\sin\theta_1}\right|^2 = \left|\frac{2(i0.3536)\frac{1}{\sqrt{2}}}{\frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}} + (i0.3536)(1.061)}\right|^2 = 0.640$$

$$r_p^{1\to2} = -\frac{\cos\theta_1\sin\theta_1 - \cos\theta_0\sin\theta_0}{\cos\theta_1\sin\theta_1 + \cos\theta_0\sin\theta_0} = -\frac{(i0.3536)(1.061) - \frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}}}{(i0.3536)(1.061) + \frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}}} = e^{-i1.287}$$

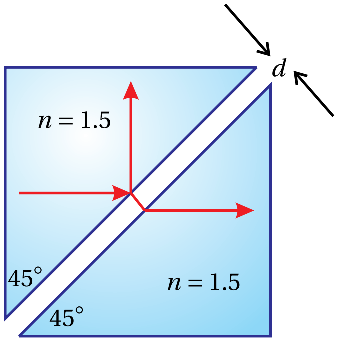


Figure 4.6 Frustrated total internal reflection in two prisms.

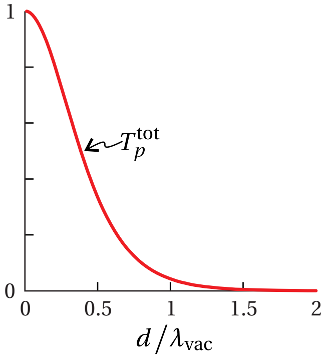


Figure 4.7 Plot of (4.22)

For the last step in the $r_p^{1\to2}$ calculation, see problem P0.15. Also note that $r_p^{1\to2} = r_p^{0\leftarrow1} = -r_p^{0\to1}$ since $n_0 = n_2$. We also need

$$k_1 d\cos\theta_1 = \frac{2\pi}{\lambda_{\text{vac}}} d\cos\theta_1 = \frac{2\pi}{\lambda_{\text{vac}}} d\,(i0.3536) = i2.22\left(\frac{d}{\lambda_{\text{vac}}}\right)$$

We are now ready to compute the total transmittance (4.14). The factors out in front vanish since $\theta_0 = \theta_2$ and $n_0 = n_2$, and we have

$$\begin{aligned}
T_p^{\text{tot}} &= \frac{\left|t_p^{0\to1}\right|^2\left|t_p^{1\to2}\right|^2}{\left|e^{-ik_1 d\cos\theta_1} - r_p^{0\leftarrow1} r_p^{1\to2} e^{ik_1 d\cos\theta_1}\right|^2} \\
&= \frac{(5.76)(0.640)}{\left|e^{-i\left[i2.22\left(\frac{d}{\lambda_{\text{vac}}}\right)\right]} - e^{-i1.287} e^{-i1.287} e^{i\left[i2.22\left(\frac{d}{\lambda_{\text{vac}}}\right)\right]}\right|^2} \\
&= \frac{3.69}{\left(e^{2.22\left(\frac{d}{\lambda_{\text{vac}}}\right)} - e^{-2.22\left(\frac{d}{\lambda_{\text{vac}}}\right)-i2.574}\right)\left(e^{2.22\left(\frac{d}{\lambda_{\text{vac}}}\right)} - e^{-2.22\left(\frac{d}{\lambda_{\text{vac}}}\right)+i2.574}\right)} \\
&= \frac{3.69}{e^{4.44\left(\frac{d}{\lambda_{\text{vac}}}\right)} + e^{-4.44\left(\frac{d}{\lambda_{\text{vac}}}\right)} - 2\left(\frac{e^{i2.574}+e^{-i2.574}}{2}\right)} \\
&= \frac{3.69}{e^{4.44\left(\frac{d}{\lambda_{\text{vac}}}\right)} + e^{-4.44\left(\frac{d}{\lambda_{\text{vac}}}\right)} - 2\cos(2.574)} \\
&= \frac{3.69}{e^{4.44\left(\frac{d}{\lambda_{\text{vac}}}\right)} + e^{-4.44\left(\frac{d}{\lambda_{\text{vac}}}\right)} + 1.69}
\end{aligned} \tag{4.22}$$

Figure 4.7 shows a plot of the transmittance (4.22) calculated in Example 4.3. Notice that the transmittance is 100% when the two prisms are brought together as expected. That is, $T_p^{\text{tot}}(d=0) = 1$. When the prisms are about a wavelength apart, the transmittance is significantly reduced, and as the distance gets large compared to a wavelength, the transmittance quickly goes to zero ($T_p^{\text{tot}}(d/\lambda_{\text{vac}} \gg 1) \approx 0$).

4.4 Fabry-Perot Instrument

In the 1890s, Charles Fabry realized that a double interface could be used to distinguish wavelengths of light that are very close together. He and a talented experimentalist colleague, Alfred Perot, constructed an instrument and began to use it to make measurements on various spectral sources. The Fabry-Perot instrument[6] consists of two identical (parallel) surfaces separated by spacing d. We can use our analysis in section 4.2 to describe this instrument. For simplicity, we choose the refractive index before the initial surface and after the final surface to be the same (i.e. $n_0 = n_2$). We assume that the transmission angles are such that total internal reflection is avoided. The transmission through the device depends on the exact spacing between the two surfaces, the reflectance of the surfaces, as well as on the wavelength of the light.

Maurice Paul Auguste Charles Fabry (1867-1945, French) was born in Marseille, France. At age 18, he entered the École Polytechnique in Paris where he studied for two years. Following that, he spent a number of years teaching state secondary school while simultaneously working on a doctoral dissertation on interference phenomena. After completing his doctorate, he began working as a lecturer and laboratory assistant at the University of Marseille where a decade later he was appointed a professor of physics. Soon after his arrival to the University of Marseille, Fabry began a long and fruitful collaboration with Alfred Perot (1863-1925). Fabry focused on theoretical analysis and measurements while his colleague did the design work and construction of their new interferometer, which they continually improved over the years. During his career, Fabry made significant contributions to spectroscopy and astrophysics and is credited with co-discovery of the ozone layer. See J. F. Mulligan, "Who were Fabry and Perot?," Am. J. Phys. **66**. 797-802 (1998).

---

[6] M. Born and E. Wolf, *Principles of Optics*, 7th ed., Sect. 7.6.2 (Cambridge University Press, 1999).

If the spacing d separating the two parallel surfaces is adjustable, the instrument is called a *Fabry-Perot interferometer*. If the spacing is fixed while the angle of the incident light is varied, the instrument is called a *Fabry-Perot etalon*. An etalon can therefore be as simple as a piece of glass with parallel surfaces. Sometimes, a thin optical membrane called a *pellicle* is used as an etalon (occasionally inserted into laser cavities to discriminate against certain wavelengths). However, to achieve sharp discrimination between closely-spaced wavelengths, a relatively large spacing d is desirable.

As we previously derived (4.15), the transmittance through a double boundary is

$$T^{\text{tot}} = \frac{T^{\max}}{1 + F\sin^2\left(\frac{\Phi}{2}\right)} \tag{4.23}$$

In the case of identical interfaces, the transmittance and reflectance coefficients are the same at each surface (i.e. $T = T^{0\to1} = T^{1\to2}$ and $R = R^{0\leftarrow1} = R^{1\to2}$). In this case, the maximum transmittance and the finesse coefficient simplify to

$$T^{\max} = \frac{T^2}{(1-R)^2} \tag{4.24}$$

and

$$F = \frac{4R}{(1-R)^2} \tag{4.25}$$

In principle, these equations should be evaluated for either s- or p-polarized light. However, a Fabry-Perot interferometer or etalon is usually operated near normal incidence so that there is little difference between the two polarizations.

When using a Fabry-Perot instrument, one observes the transmittance T^{tot} as the parameter Φ is varied. The parameter Φ can be varied by altering d, θ_1, or λ as prescribed by

$$\Phi = \frac{4\pi n_1 d}{\lambda_{\text{vac}}}\cos\theta_1 + 2\phi_r \tag{4.26}$$

To increase the sensitivity of the instrument, it is desirable to have the transmittance T^{tot} vary strongly as a function of Φ. By inspection of (4.23), we see that this happens if the finesse coefficient F is large. We achieve a large finesse coefficient by increasing the reflectance R.

The basic Fabry-Perot instrument design is shown in Fig. 4.8. In order to achieve high reflectivity R (and therefore large F), special coatings can be applied to the surfaces, for example, a thin layer of silver to achieve reflectance of, say, 90%. Typically, two glass substrates are separated by distance d, with the coated surfaces facing each other as shown in the figure. The substrates are aligned so that the interior surfaces are parallel to each other. It is typical for each substrate to be slightly wedge-shaped so that unwanted reflections from the outer surfaces do not interfere with the double boundary situation between the two plates.

Technically, each coating constitutes its own double-boundary problem (or multiple-boundary as the case may be). We can ignore this detail and simply think of the overall setup as a single two-interface problem. Regardless of the

Jean-Baptiste Alfred Perot (1863-1925, French) was born in Metz, France. He attended the Ecole Polytechnique and then the University of Paris, where he earned a doctorate in 1888. He became a professor in Marseille in 1894 where he began his collaboration with Fabry. Perot contributed his considerable talent of instrument fabrication to the endeavor. Perot spent much of his later career making precision astronomical and solar measurements. See J. F. Mulligan, "Who were Fabry and Perot?," Am. J. Phys. **66**. 797-802 (1998).

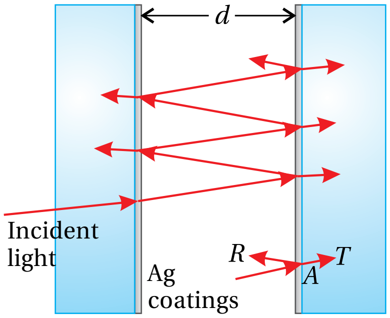


Figure 4.8 Typical Fabry-Perot setup. If the spacing d is variable, it is called an interferometer; otherwise, it is called an etalon.

details of the coatings, we can say that each coating has a certain reflectance R and transmittance T. However, as light goes through a coating, it can also be attenuated because of absorption. In this case, we have

$$R + T + A = 1 \tag{4.27}$$

where A represents the amount of light absorbed at a coating. The attenuation A reduces the amount of light that makes it through the instrument, but it does not impact the nature of the interferences within the instrument.

The total transmittance T^{tot} (4.23) through an ideal Fabry-Perot instrument is depicted in Fig. 4.9 as a function of Φ. The various curves correspond to different values of F. Typical values of Φ can be extremely large. For example, suppose that the instrument is used at near-normal incidence (i.e. $\cos\theta_1 \cong 1$) with a wavelength of $\lambda_{\text{vac}} = 500$ nm and an interface separation of $d = 1$ cm. From (4.26) the value of Φ (ignoring the phase term $2\phi_r$) is approximately

$$\Phi = \frac{4\pi(1\text{ cm})}{500\text{ nm}} = 80,000\pi$$

As we vary d, λ, or θ_1 by small amounts, we can easily cause Φ to change by 2π as depicted in Fig. 4.9. The figure shows small changes in Φ in the neighborhood of very large multiples of 2π.

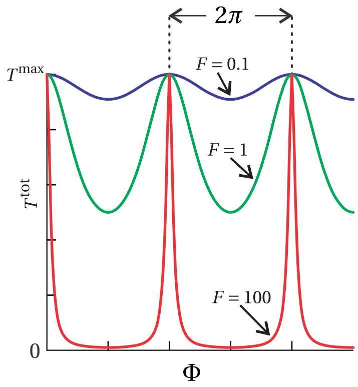


Figure 4.9 Transmittance as the phase Φ is varied. The different curves correspond to different values of the finesse coefficient.

The phase term $2\phi_r$ in (4.26) depends on the exact nature of the coatings in the Fabry-Perot instrument. However, we do not need to know the value of ϕ_r, which may depend on both the complex index of the coating material and its thickness. Whatever the value of ϕ_r, we only care that it is constant. Experimentally, we can always compensate for the ϕ_r by 'tweaking' the spacing d, whose exact value is likely not controlled for in the first place. Note that the required 'tweak' on the spacing need only be a fraction of a wavelength, which is typically tiny compared to the overall spacing d.

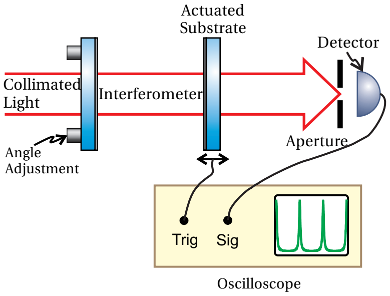


Figure 4.10 Setup for a Fabry-Perot interferometer.

4.5 Setup of a Fabry-Perot Instrument

Figure 4.10 shows the typical experimental setup for a *Fabry-Perot interferometer*. A *collimated* beam of light is sent through the instrument. The beam is aligned so that it is normal to the surfaces. It is critical for the two surfaces of the interferometer to be extremely close to parallel. When aligned correctly, the transmission of a collimated beam will 'blink' all together as the spacing d is changed (by tiny amounts). A mechanical actuator can be used to vary the spacing between the plates while the transmittance is observed on a detector. To make the alignment of the instrument somewhat less critical, a small aperture can be placed in front of the detector so that it observes only a small portion of the beam.

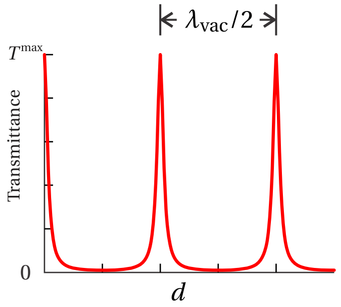


Figure 4.11 Transmittance as the separation d is varied ($F = 100$).

The transmittance as a function of plate separation is shown in Fig. 4.11. In this case, Φ varies via changes in d (see (4.26) with $\cos\theta_1 = 1$ and fixed wavelength). As the spacing is increased by only a half wavelength, the transmittance changes through a complete period. The various peaks in the figure are called *fringes*.

The setup for a *Fabry-Perot etalon* is similar to that of the interferometer except that the spacing d remains fixed. Often the two surfaces in the etalon are held parallel to each other by a precision spacer. An advantage to the Fabry-Perot etalon (as opposed to the interferometer) is that no moving parts are needed. To make measurements with an etalon, the angle of the light is varied rather than the plate separation. After all, to see fringes, we just need to cause Φ in (4.23) to vary in some way. According to (4.26), we can do that as easily by varying θ_1 as we can by varying d. One way to obtain a range of angles is to observe light from a 'point source', as depicted in Fig. 4.12. Different portions of the beam go through the device at different angles. When aligned straight on, the transmitted light forms a 'bull's-eye' pattern on a screen.

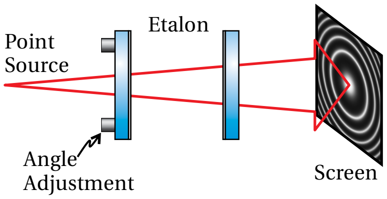


Figure 4.12 Schematic of a diverging monochromatic beam traversing a Fabry-Perot etalon. The divergence angle is exaggerated.

In Fig. 4.13 we graph the transmittance T^{tot} (4.23) as a function of angle (holding $\lambda_{\text{vac}} = 500$ nm and $d = 1$ cm fixed). Since $\cos\theta_1$ is not a linear function, the spacing of the peaks varies with angle. As θ_1 increases from zero, the cosine steadily decreases, causing Φ to decrease. Each time Φ decreases by 2π we get a new peak. Not surprisingly, only a modest change in angle is necessary to cause the transmittance to vary from maximum to minimum, or vice versa.

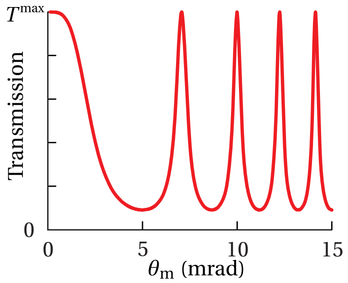


Figure 4.13 Transmittance through a Fabry-Perot etalon ($F = 10$) as the angle θ_1 is varied. It is assumed that the distance d is chosen such that Φ is a multiple of 2π when the angle is zero.

The bull's-eye pattern in Fig. 4.12 can be understood as the curve in Fig. 4.13 rotated about a circle. Depending on the exact spacing between the plates, the angles where the fringes occur can be different. For example, the center spot could be dark.

Spectroscopic samples often are not compact point-like sources. Rather, they are extended diffuse sources. The point-source setup shown in Fig. 4.12 won't work for extended sources unless all of the light at the sample is blocked except for a tiny point. This is impractical if there remains insufficient illumination at the final screen for observation.

In order to preserve as much light as possible, we can sandwich the etalon between two lenses. We place the diffuse source at the focal plane of the first lens. We place the screen at the focal plane of the second lens. This causes an image of the source to appear on the screen.[7] Each point of the diffuse source is mapped to a corresponding point on the screen. Moreover, the light associated with any particular point of the source travels as a unique collimated beam in the region between the lenses. Each collimated beam traverses the etalon with a specific angle. Thus, light associated with each emission point traverses the etalon with higher or lower transmittance, according to the differing angles. The result is that a bull's eye pattern becomes superimposed on the image of the diffuse source. The lens and retina of your eye can be used for the final lens and screen.

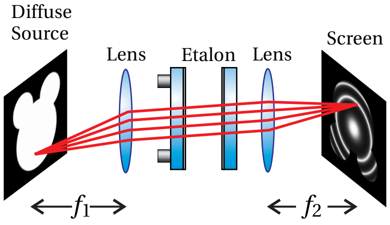


Figure 4.14 Setup of a Fabry-Perot etalon for looking at a diffuse source.

---

[7] If the diffuse source has the shape of Mickey Mouse, then an image of Mickey Mouse appears on the screen. Imaging techniques are discussed in chapter 9.

4.6 Distinguishing Nearby Wavelengths in a Fabry-Perot Instrument

Thus far, we have examined how the transmittance through a Fabry-Perot instrument varies with surface separation d and angle θ_1. However, the main purpose of a Fabry-Perot instrument is to measure small changes in the wavelength of light, which similarly affect the value of Φ (see (4.26)).[8]

Consider a Fabry-Perot interferometer where the transmittance through the instrument is plotted as a function of plate spacing d. At certain spacings, Φ happens to be a multiple of 2π for the wavelength λ_{vac}. Next, suppose we adjust the wavelength to $\lambda_{\text{vac}} + \Delta\lambda$ while observing the locations of these fringes. As the wavelength changes, the locations at which Φ is a multiple of 2π change. Consequently, the fringes shift as seen in figure 4.15.

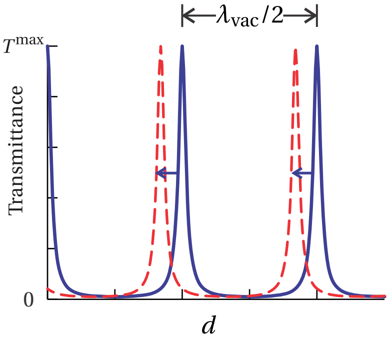


Figure 4.15 Transmittance as the spacing d is varied for two different wavelengths ($F = 100$). The solid line plots the transmittance of light with a wavelength of λ_{vac}, and the dashed line plots the transmittance of a wavelength shorter than λ_{vac}. Note that the fringes shift positions for different wavelengths.

We now derive the connection between a change in wavelength and the amount that Φ changes, which gives rise to the fringe shift seen in Fig. 4.15. At the wavelength $\lambda_{\text{vac}} + \Delta\lambda$ (all else remaining the same), (4.26) shifts to

$$\Phi - \Delta\Phi = \frac{4\pi n_1 d \cos\theta_1}{\lambda_{\text{vac}} + \Delta\lambda} + 2\phi_{\text{r}} \tag{4.28}$$

The change in wavelength $\Delta\lambda$ is usually very small compared to λ_{vac}, so we can represent the denominator with a truncated Taylor-series expansion:

$$\frac{1}{\lambda_{\text{vac}} + \Delta\lambda} = \frac{1}{\lambda_{\text{vac}}\left(1 + \Delta\lambda/\lambda_{\text{vac}}\right)} \cong \frac{1 - \Delta\lambda/\lambda_{\text{vac}}}{\lambda_{\text{vac}}} \tag{4.29}$$

The amount that Φ changes is then seen to be

$$\Delta\Phi = \frac{4\pi n_1 d \cos\theta_1}{\lambda_{\text{vac}}^2}\Delta\lambda \tag{4.30}$$

If the change in wavelength is enough to cause $\Delta\Phi = 2\pi$, the fringes in Fig. 4.15 shift through a whole period, and the picture looks the same.

This brings up an important limitation of the instrument. If the fringes shift by too much, we might become confused as to what exactly has changed, owing to the periodic nature of the fringes. If two wavelengths aren't sufficiently close, the fringes of one wavelength may be shifted past several fringes of the other wavelength, and we will not be able to tell by how much they differ.

This introduces the concept of *free spectral range*, which is the wavelength change $\Delta\lambda_{\text{FSR}}$ that causes the fringes to shift through one period. We find this by setting (4.30) equal to 2π. After rearranging, we get

$$\Delta\lambda_{\text{FSR}} = \frac{\lambda_{\text{vac}}^2}{2 n_1 d \cos\theta_1} \tag{4.31}$$

The free spectral range tends to be extremely narrow; a Fabry-Perot instrument is not well suited for measuring wavelength ranges wider than this. In summary, the free spectral range is the largest change in wavelength permissible while avoiding confusion. To convert this wavelength difference $\Delta\lambda_{\text{FSR}}$ into a corresponding frequency difference, one differentiates $\nu = c/\lambda_{\text{vac}}$ to get

---

[8] M. Born and E. Wolf, *Principles of Optics*, 7th ed., Sect. 7.6.3 (Cambridge University Press, 1999).

$$|\Delta\nu_{\text{FSR}}| = \frac{c\Delta\lambda_{\text{FSR}}}{\lambda_{\text{vac}}^2} \tag{4.32}$$

Example 4.4

A Fabry-Perot interferometer has plate spacing $d = 1$ cm and index $n_1 = 1$. If it is used in the neighborhood of $\lambda_{\text{vac}} = 500$ nm, find the free spectral range of the instrument.

Solution: From (4.31), the free spectral range is

$$\Delta\lambda_{\text{FSR}} = \frac{\lambda_{\text{vac}}^2}{2n_1 d_0 \cos\theta_1} = \Delta\lambda_{\text{FSR}} = \frac{(500 \text{ nm})^2}{2\,(1)\,(1 \text{ cm}) \cos 0°} = 0.0125 \text{ nm}$$

This means that we should not use the instrument to distinguish wavelengths that are separated by more than this small amount.

We next consider the *smallest* change in wavelength that can be noticed, or *resolved* with a Fabry-Perot instrument. For example, if two very near-by wavelengths are sent through the instrument simultaneously, we can distinguish them only if the separation between their corresponding fringe peaks is at least as large as the width of an individual peak. This situation of two barely resolvable fringe peaks is illustrated in Fig. 4.16 for a diverging beam traversing an etalon.

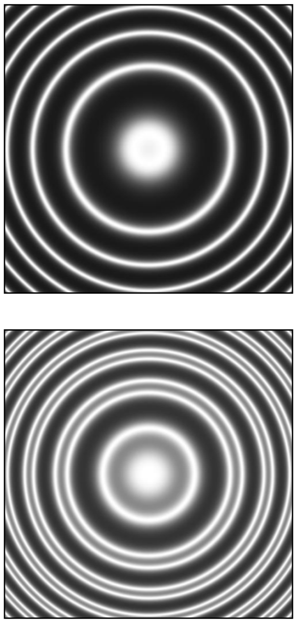


Figure 4.16 Transmittance of a diverging beam through a Fabry-Perot etalon. Two nearby wavelengths are sent through the instrument simultaneously, (top) barely resolved and (bottom) easily resolved.

We will look for the wavelength change that causes a peak to shift by its own width. We define the width of a peak by its *full width at half maximum* (FWHM). Again, let Φ be a multiple of 2π where a peak in transmittance occurs. In this case, we have from (4.23) that

$$T^{\text{tot}} = \frac{T^{\text{max}}}{1 + F\sin^2\left(\frac{\Phi}{2}\right)} = T^{\text{max}} \tag{4.33}$$

since $\sin(\Phi/2) = 0$. When Φ shifts to a neighboring value $\Phi \pm \Phi_{\text{FWHM}}/2$, then, by definition, the transmittance drops to one half. Therefore, we may write

$$T^{\text{tot}} = \frac{T^{\text{max}}}{1 + F\sin^2\left(\frac{\Phi_0 \pm \Phi_{\text{FWHM}}/2}{2}\right)} = \frac{T^{\text{max}}}{2} \tag{4.34}$$

In solving for (4.34) for Φ_{FWHM}, we see that this equation requires

$$F\sin^2\left(\frac{\Phi_{\text{FWHM}}}{4}\right) = 1 \tag{4.35}$$

where we have taken advantage of the fact that Φ is assumed to be a multiple of 2π. Next, we suppose that $\Phi_{\text{FWHM}}/4$ is rather small so that we may represent the sine by its argument. This approximation is okay if the finesse coefficient F is rather large (say, 100). With this approximation, (4.35) simplifies to

$$\Phi_{\text{FWHM}} \cong \frac{4}{\sqrt{F}}. \tag{4.36}$$

The ratio of the period between peaks 2π to the width Φ_{FWHM} of individual peaks is called the *reflecting finesse* (or just *finesse*).

$$f \equiv \frac{2\pi}{\Phi_{\text{FWHM}}} = \frac{\pi\sqrt{F}}{2} \tag{4.37}$$

This parameter is often used to characterize the performance of a Fabry-Perot instrument. Note that a higher finesse f implies sharper fringes in comparison to the fringe spacing.

The free spectral range $\Delta\lambda_{\text{FSR}}$ compared to the minimum wavelength $\Delta\lambda_{\text{FWHM}}$ is the same ratio f. Therefore, we have

$$\Delta\lambda_{\text{FWHM}} = \frac{\Delta\lambda_{\text{FSR}}}{f} = \frac{\lambda_{\text{vac}}^2}{\pi n_1 d \cos\theta_1 \sqrt{F}} \tag{4.38}$$

As a final note, the ratio of λ_{vac} to $\Delta\lambda_{\text{FWHM}}$, where $\Delta\lambda_{\text{FWHM}}$ is the minimum change of wavelength that the instrument can distinguish in the neighborhood of λ_{vac}, is called the *resolving power*:

$$\text{RP} \equiv \frac{\lambda_{\text{vac}}}{\Delta\lambda_{\text{FWHM}}} \tag{4.39}$$

Fabry-Perot instruments tend to have very high resolving powers as the following example illustrates.

Example 4.5

If the Fabry-Perot interferometer in Example 4.4 has reflectivity $R = 0.85$, find the finesse, the minimum distinguishable wavelength separation, and the resolving power.

Solution: From (4.25), the finesse coefficient is

$$F = \frac{4R}{(1-R)^2} = \frac{4\,(0.85)}{(1-(0.85))^2} = 151$$

and by (4.37) the finesse is

$$f = \frac{\pi\sqrt{F}}{2} = \frac{\pi\sqrt{151}}{2} = 19.3$$

The minimum resolvable wavelength change is then

$$\Delta\lambda_{\text{FWHM}} = \frac{\Delta\lambda_{\text{FSR}}}{f} = \frac{0.0125\ \text{nm}}{19} = 0.00065\ \text{nm} \tag{4.40}$$

> The instrument can distinguish two wavelengths separated by this tiny amount, which gives an impressive resolving power of
>
> $$\mathrm{RP} = \frac{\lambda_{\mathrm{vac}}}{\Delta\lambda_{\mathrm{FWHM}}} = \frac{500\ \mathrm{nm}}{0.00065\ \mathrm{nm}} = 772{,}000$$
>
> For comparison, the resolving power of a typical grating spectrometer is much less (a few thousand). However, a grating spectrometer has the advantage that it can simultaneously observe wavelengths over hundreds of nanometers, whereas the Fabry-Perot instrument is confined to the extremely narrow free spectral range.

4.7 Multilayer Coatings

As we saw in Example 4.2, a single coating cannot always accomplish a desired effect, especially if the goal is to make a highly reflective mirror. For example, if we want to make a mirror surface using a dielectric (i.e. nonmetallic) coating, a single layer is insufficient to reflect the majority of the light. In P4.5 we find that a single dielectric layer deposited on glass can reflect at most about 46% of the light, even when we used a material with very high index. We would like to do much better (e.g. >99%), and this can be accomplished with multilayer dielectric coatings. Multilayer dielectric coatings can perform considerably better than metal surfaces such as silver and have the advantage of being less prone to damage.

In this section, we develop the formalism for dealing with arbitrary numbers of parallel interfaces (i.e. multilayer coatings).[9] Rather than incorporate the single-interface Fresnel coefficients into the problem as we did in section 4.1, we will find it easier to return to the fundamental boundary conditions for the electric and magnetic fields at each interface between the layers.

We examine p-polarized light incident on an arbitrary multilayer coating with all interfaces parallel to each other. It is left as an exercise to re-derive the formalism for s-polarized light (see P4.13). The upcoming derivation is valid also for complex refractive indices, although our notation suggests real indices. The ability to deal with complex indices is very important if, for example, we want to make mirror coatings work in the extreme ultraviolet wavelength range where virtually every material is absorptive. Consider the diagram of a multilayer coating in Fig. 4.17 for which the angle of light propagation in each region may be computed from Snell's law:

$$n_0 \sin\theta_0 = n_1 \sin\theta_1 = \cdots = n_N \sin\theta_N = n_{N+1} \sin\theta_{N+1} \tag{4.41}$$

where N denotes the number of layers in the coating. The subscript 0 represents the initial medium outside of the multilayer, and the subscript $N+1$ represents the final material, or the substrate on which the layers are deposited.

---

[9] G. R. Fowles, *Introduction to Modern Optics*, 2nd ed., Sect 4.4 (New York: Dover, 1975); E. Hecht, *Optics*, 3rd ed., Sect. 9.7.1 (Massachusetts: Addison-Wesley, 1998).

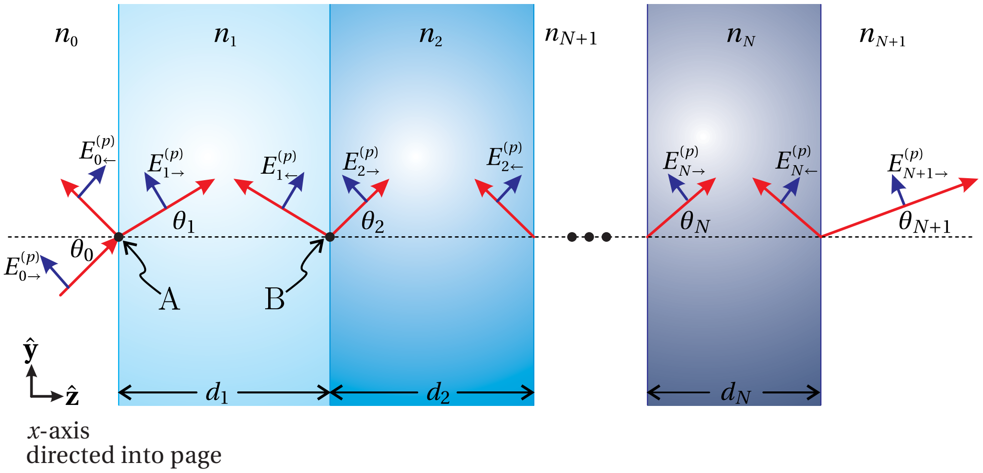

n_0
n_1
n_2
n_{N+1}
n_N
n_{N+1}
$E^{(p)}_{0\leftarrow}$
$E^{(p)}_{1\to}$
$E^{(p)}_{1\leftarrow}$
$E^{(p)}_{2\to}$
$E^{(p)}_{2\leftarrow}$
$E^{(p)}_{N\to}$
$E^{(p)}_{N\leftarrow}$
$E^{(p)}_{N+1\to}$
θ_1
θ_2
θ_N
θ_{N+1}
θ_0
$E^{(p)}_{0\to}$
A
B
$\hat{\mathbf{y}}$
$\hat{\mathbf{z}}$
d_1
d_2
d_N
x-axis
directed into page

Figure 4.17 Light propagation through multiple layers.

In each layer, only two plane waves exist, each of which is composed of light arising from the many possible bounces from various layer interfaces. The arrows pointing right indicate plane wave fields in individual layers that travel roughly in the forward (incident) direction, and the arrows pointing left indicate plane wave fields that travel roughly in the backward (reflected) direction. In the final region, there is only one plane wave traveling with a forward direction ($E^{(p)}_{N+1\to}$) which gives the overall transmitted field.

As we have studied in chapter 3 (see (3.9) and (3.13)), the boundary conditions for the parallel components of the **E** field and for the parallel components of the **B** field lead respectively to

$$\cos\theta_0\left(E^{(p)}_{0\to}+E^{(p)}_{0\leftarrow}\right)=\cos\theta_1\left(E^{(p)}_{1\to}+E^{(p)}_{1\leftarrow}\right) \tag{4.42}$$

and

$$n_0\left(E^{(p)}_{0\to}-E^{(p)}_{0\leftarrow}\right)=n_1\left(E^{(p)}_{1\to}-E^{(p)}_{1\leftarrow}\right) \tag{4.43}$$

Similar equations give the field connection for s-polarized light (see (3.8) and (3.14)).

We have applied these boundary conditions at the first interface only. Of course there are many more interfaces in the multilayer. For the connection between the j^{th} layer and the next, we may similarly write

$$\cos\theta_j\left(E^{(p)}_{j\to}e^{ik_jd_j\cos\theta_j}+E^{(p)}_{j\leftarrow}e^{-ik_jd_j\cos\theta_j}\right)=\cos\theta_{j+1}\left(E^{(p)}_{j+1\to}+E^{(p)}_{j+1\leftarrow}\right) \tag{4.44}$$

and

$$n_j\left(E^{(p)}_{j\to}e^{ik_jd_j\cos\theta_j}-E^{(p)}_{j\leftarrow}e^{-ik_jd_j\cos\theta_j}\right)=n_{j+1}\left(E^{(p)}_{j+1\to}-E^{(p)}_{j+1\leftarrow}\right) \tag{4.45}$$

Here we have set the origin within each layer at the left surface. Then when making the connection with the subsequent layer at the right surface, we must specifically take into account the phase $\mathbf{k}_j\cdot(d_j\hat{\mathbf{z}})=k_jd_j\cos\theta_j$. This corresponds to the phase acquired by the plane wave field in traversing the layer with thickness d_j. The right-hand sides of (4.44) and (4.45) need no phase adjustment since the $(j+1)^{\text{th}}$ field is evaluated on the left side of its layer.

At the final interface, the boundary conditions are

$$\cos\theta_N \left(E^{(p)}_{N\rightarrow} e^{ik_N d_N \cos\theta_N} + E^{(p)}_{N\leftarrow} e^{-ik_N d_N \cos\theta_N} \right) = \cos\theta_{N+1} E^{(p)}_{N+1\rightarrow} \tag{4.46}$$

and

$$n_N \left(E^{(p)}_{N\rightarrow} e^{ik_N d_N \cos\theta_N} - E^{(p)}_{N\leftarrow} e^{-ik_N d_N \cos\theta_N} \right) = n_{N+1} E^{(p)}_{N+1\rightarrow} \tag{4.47}$$

since there is no backward-traveling field in the final medium.

At this point we are ready to solve (4.42)–(4.47). We would like to eliminate all fields besides $E^{(p)}_{0\rightarrow}$, $E^{(p)}_{0\leftarrow}$, and $E^{(p)}_{N+1\rightarrow}$. Then we will be able to find the overall reflectance and transmittance of the multilayer coating. In solving (4.42)–(4.47), we must proceed with care, or the algebra can quickly get out of hand. Fortunately, you have probably had training in linear algebra, and this is a case where that training pays off.

We first write a general matrix equation that summarizes the mathematics in (4.42)–(4.47), as follows:

$$\begin{bmatrix} \cos\theta_j e^{i\beta_j} & \cos\theta_j e^{-i\beta_j} \\ n_j e^{i\beta_j} & -n_j e^{-i\beta_j} \end{bmatrix} \begin{bmatrix} E^{(p)}_{j\rightarrow} \\ E^{(p)}_{j\leftarrow} \end{bmatrix} = \begin{bmatrix} \cos\theta_{j+1} & \cos\theta_{j+1} \\ n_{j+1} & -n_{j+1} \end{bmatrix} \begin{bmatrix} E^{(p)}_{j+1\rightarrow} \\ E^{(p)}_{j+1\leftarrow} \end{bmatrix} \tag{4.48}$$

where

$$\beta_j \equiv \begin{cases} 0 & j = 0 \\ k_j d_j \cos\theta_j & 1 \le j \le N \end{cases} \tag{4.49}$$

and

$$E^{(p)}_{N+1\leftarrow} \equiv 0 \tag{4.50}$$

(It would be good to take a moment to convince yourself that this set of matrix equations properly represents (4.42)–(4.47) before proceeding.) We rewrite (4.48) as

$$\begin{bmatrix} E^{(p)}_{j\rightarrow} \\ E^{(p)}_{j\leftarrow} \end{bmatrix} = \begin{bmatrix} \cos\theta_j e^{i\beta_j} & \cos\theta_j e^{-i\beta_j} \\ n_j e^{i\beta_j} & -n_j e^{-i\beta_j} \end{bmatrix}^{-1} \begin{bmatrix} \cos\theta_{j+1} & \cos\theta_{j+1} \\ n_{j+1} & -n_{j+1} \end{bmatrix} \begin{bmatrix} E^{(p)}_{j+1\rightarrow} \\ E^{(p)}_{j+1\leftarrow} \end{bmatrix} \tag{4.51}$$

Keep in mind that (4.51) represents a distinct matrix equation for each different j. We can substitute the $j = 1$ equation into the $j = 0$ equation to get

$$\begin{bmatrix} E^{(p)}_{0\rightarrow} \\ E^{(p)}_{0\leftarrow} \end{bmatrix} = \begin{bmatrix} \cos\theta_0 & \cos\theta_0 \\ n_0 & -n_0 \end{bmatrix}^{-1} M^{(p)}_1 \begin{bmatrix} \cos\theta_2 & \cos\theta_2 \\ n_2 & -n_2 \end{bmatrix} \begin{bmatrix} E^{(p)}_{2\rightarrow} \\ E^{(p)}_{2\leftarrow} \end{bmatrix} \tag{4.52}$$

where we have grouped the matrices related to the $j = 1$ layer together via

$$M^{(p)}_1 \equiv \begin{bmatrix} \cos\theta_1 & \cos\theta_1 \\ n_1 & -n_1 \end{bmatrix} \begin{bmatrix} \cos\theta_1 e^{i\beta_1} & \cos\theta_1 e^{-i\beta_1} \\ n_1 e^{i\beta_1} & -n_1 e^{-i\beta_1} \end{bmatrix}^{-1} \tag{4.53}$$

We can continue to substitute into this equation progressively higher order equations (i.e. for $j = 2$, $j = 3$, ...) until we reach the $j = N$ layer. All together this will give

$$\begin{bmatrix} E^{(p)}_{0\to} \\ E^{(p)}_{0\leftarrow} \end{bmatrix} = \begin{bmatrix} \cos\theta_0 & \cos\theta_0 \\ n_0 & -n_0 \end{bmatrix}^{-1} \left(\prod_{j=1}^{N} M_j^{(p)} \right) \begin{bmatrix} \cos\theta_{N+1} & \cos\theta_{N+1} \\ n_{N+1} & -n_{N+1} \end{bmatrix} \begin{bmatrix} E^{(p)}_{N+1\to} \\ 0 \end{bmatrix} \tag{4.54}$$

where the matrices related to the j^{th} layer are grouped together according to

$$\begin{aligned} M_j^{(p)} &\equiv \begin{bmatrix} \cos\theta_j & \cos\theta_j \\ n_j & -n_j \end{bmatrix} \begin{bmatrix} \cos\theta_j e^{i\beta_j} & \cos\theta_j e^{-i\beta_j} \\ n_j e^{i\beta_j} & -n_j e^{-i\beta_j} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \cos\beta_j & -i\sin\beta_j \cos\theta_j / n_j \\ -i n_j \sin\beta_j / \cos\theta_j & \cos\beta_j \end{bmatrix} \end{aligned} \tag{4.55}$$

The matrix inversion in the first line was performed using (0.35). The symbol Π signifies the product of the matrices with the lowest subscripts on the left:

$$\prod_{j=1}^{N} M_j^{(p)} \equiv M_1^{(p)} M_2^{(p)} \cdots M_N^{(p)} \tag{4.56}$$

As a finishing touch, we divide (4.54) by the incident field $E^{(p)}_{0\to}$ as well as perform the matrix inversion on the right-hand side to obtain

$$\begin{bmatrix} 1 \\ E^{(p)}_{0\leftarrow} \big/ E^{(p)}_{0\to} \end{bmatrix} = A^{(p)} \begin{bmatrix} E^{(p)}_{N+1\to} \big/ E^{(p)}_{0\to} \\ 0 \end{bmatrix} \tag{4.57}$$

where

$$A^{(p)} \equiv \begin{bmatrix} a^{(p)}_{11} & a^{(p)}_{12} \\ a^{(p)}_{21} & a^{(p)}_{22} \end{bmatrix} = \frac{1}{2n_0\cos\theta_0} \begin{bmatrix} n_0 & \cos\theta_0 \\ n_0 & -\cos\theta_0 \end{bmatrix} \left(\prod_{j=1}^{N} M_j^{(p)} \right) \begin{bmatrix} \cos\theta_{N+1} & 0 \\ n_{N+1} & 0 \end{bmatrix} \tag{4.58}$$

In the final matrix in (4.58) we have replaced the entries in the right column with zeros. This is permissible since it operates on a column vector with zero in the bottom component.

Equation (4.57) represents two equations, which must be solved simultaneously to find the ratios $E^{(p)}_{0\leftarrow} / E^{(p)}_{0\to}$ and $E^{(p)}_{N+1\to} / E^{(p)}_{0\to}$. Once the matrix $A^{(p)}$ is computed, this is a relatively simple task:

$$t_p^{\text{tot}} \equiv \frac{E^{(p)}_{N+1\to}}{E^{(p)}_{0\to}} = \frac{1}{a^{(p)}_{11}} \qquad \text{(Multilayer)} \tag{4.59}$$

$$r_p^{\text{tot}} \equiv \frac{E^{(p)}_{0\leftarrow}}{E^{(p)}_{0\to}} = \frac{a^{(p)}_{21}}{a^{(p)}_{11}} \qquad \text{(Multilayer)} \tag{4.60}$$

The convenience of this notation lies in the fact that we can deal with an arbitrary number of layers N with varying thickness and index. The essential information for each layer is contained succinctly in its respective 2 × 2 *characteristic matrix* M. To find the overall effect of the many layers, we need only multiply the matrices for each layer together to find A from which we compute the reflection and transmission coefficients for the whole system.

The derivation for s-polarized light is similar to the above derivation for p-polarized light. The equation corresponding to (4.57) for s-polarized light turns out to be

$$\begin{bmatrix} 1 \\ E_{0\leftarrow}^{(s)} \big/ E_{0\rightarrow}^{(s)} \end{bmatrix} = A^{(s)} \begin{bmatrix} E_{N+1\rightarrow}^{(s)} \big/ E_{0\rightarrow}^{(s)} \\ 0 \end{bmatrix} \tag{4.61}$$

where

$$A^{(s)} \equiv \begin{bmatrix} a_{11}^{(s)} & a_{12}^{(s)} \\ a_{21}^{(s)} & a_{22}^{(s)} \end{bmatrix} = \frac{1}{2n_0\cos\theta_0} \begin{bmatrix} n_0\cos\theta_0 & 1 \\ n_0\cos\theta_0 & -1 \end{bmatrix} \left(\prod_{j=1}^{N} M_j^{(s)} \right) \begin{bmatrix} 1 & 0 \\ n_{N+1}\cos\theta_{N+1} & 0 \end{bmatrix} \tag{4.62}$$

and

$$M_j^{(s)} = \begin{bmatrix} \cos\beta_j & -i\sin\beta_j/(n_j\cos\theta_j) \\ -in_j\cos\theta_j\sin\beta_j & \cos\beta_j \end{bmatrix} \tag{4.63}$$

The transmission and reflection coefficients are found (as before) from

$$t_s^{\text{tot}} \equiv \frac{E_{N+1\rightarrow}^{(s)}}{E_{0\rightarrow}^{(s)}} = \frac{1}{a_{11}^{(s)}} \qquad \text{(Multilayer)} \tag{4.64}$$

$$r_s^{\text{tot}} \equiv \frac{E_{0\leftarrow}^{(s)}}{E_{0\rightarrow}^{(s)}} = \frac{a_{21}^{(s)}}{a_{11}^{(s)}} \qquad \text{(Multilayer)} \tag{4.65}$$

4.8 Periodic Multilayer Stacks

Many different types of multilayer coatings are possible. For example, a Brewster's-angle polarizer has a coating designed to transmit with high efficiency p-polarized light while simultaneously reflecting s-polarized light with high efficiency. The backside of the substrate is left uncoated where p-polarized light passes with 100% efficiency at Brewster's angle.

Sometimes multilayer coatings are made with repeated stacks of layers. If the same series of layers in (4.66) is repeated many times, say q times, Sylvester's theorem (see section 0.3) can come in handy. A block of matrices, corresponding to a repeated pattern within the stack, can be conveniently taken to any power. Sylvester's theorem requires that the determinant of the matrix be to equal one, which is true for matrices of the form (4.55) and (4.63) or any product of them.

It is common for high-reflection coatings to be designed with alternating high and low refractive indices. For high reflectivity, each layer should have a quarter-wave thickness. Since the layers alternate high and low indices, at every other boundary there is a phase shift of π upon reflection from the interface. Hence, the quarter wavelength spacing is appropriate to give constructive interference in the reflected direction.

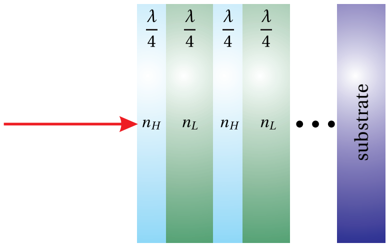


Figure 4.18 A repeated multilayer structure with alternating high and low indexes where each layer is a quarter wavelength in thickness. This structure can achieve very high reflectance.

Example 4.6

Derive the reflection and transmission coefficients for p polarized light interacting with a high reflector constructed using a $\lambda/4$ stack.

Solution: For a $\lambda/4$ stack we need

$$\beta_j = \frac{\pi}{2}$$

This amounts to a thickness requirement of

$$d_j = \frac{\lambda_{\text{vac}}}{4 n_j \cos\theta_j}$$

In this situation, the matrix (4.55) for each layer simplifies to

$$M_j^{(p)} = \begin{bmatrix} 0 & -i\cos\theta_j / n_j \\ -i n_j / \cos\theta_j & 0 \end{bmatrix}$$

The matrices for a high and a low refractive index layer are multiplied together in the usual manner. Each layer pair takes the form

$$\begin{bmatrix} 0 & -\frac{i\cos\theta_H}{n_H} \\ -\frac{i n_H}{\cos\theta_H} & 0 \end{bmatrix} \begin{bmatrix} 0 & -\frac{i\cos\theta_L}{n_L} \\ -\frac{i n_L}{\cos\theta_L} & 0 \end{bmatrix} = \begin{bmatrix} -\frac{n_L\cos\theta_H}{n_H\cos\theta_L} & 0 \\ 0 & -\frac{n_H\cos\theta_L}{n_L\cos\theta_H} \end{bmatrix}$$

To extend to $q = N/2$ identical layer pairs, we have

$$\begin{aligned} \prod_{j=1}^{N} M_j^{(p)} &= \begin{bmatrix} -\frac{n_L\cos\theta_H}{n_H\cos\theta_L} & 0 \\ 0 & -\frac{n_H\cos\theta_L}{n_L\cos\theta_H} \end{bmatrix}^q \\ &= \begin{bmatrix} \left(-\frac{n_L\cos\theta_H}{n_H\cos\theta_L}\right)^q & 0 \\ 0 & \left(-\frac{n_H\cos\theta_L}{n_L\cos\theta_H}\right)^q \end{bmatrix} \end{aligned}$$

Substituting this into (4.58), we obtain

$$A^{(p)} = \frac{1}{2}\begin{bmatrix} \left(-\frac{n_L\cos\theta_H}{n_H\cos\theta_L}\right)^q \frac{\cos\theta_{N+1}}{\cos\theta_0} + \left(-\frac{n_H\cos\theta_L}{n_L\cos\theta_H}\right)^q \frac{n_{N+1}}{n_0} & 0 \\ \left(-\frac{n_L\cos\theta_H}{n_H\cos\theta_L}\right)^q \frac{\cos\theta_{N+1}}{\cos\theta_0} - \left(-\frac{n_H\cos\theta_L}{n_L\cos\theta_H}\right)^q \frac{n_{N+1}}{n_0} & 0 \end{bmatrix} \tag{4.66}$$

With $A^{(p)}$ in hand, we can now calculate the transmission coefficient from (4.59)

$$t_p^{\text{tot}} = \frac{1}{\left(-\frac{n_L\cos\theta_H}{n_H\cos\theta_L}\right)^q \frac{\cos\theta_{N+1}}{\cos\theta_0} + \left(-\frac{n_H\cos\theta_L}{n_L\cos\theta_H}\right)^q \frac{n_{N+1}}{n_0}} \quad (\lambda/4 \text{ stack, } p\text{-polarized}) \tag{4.67}$$

and the reflection coefficient from (4.60)

$$r_p^{\text{tot}} = \frac{\left(-\frac{n_L\cos\theta_H}{n_H\cos\theta_L}\right)^q \frac{\cos\theta_{N+1}}{\cos\theta_0} - \left(-\frac{n_H\cos\theta_L}{n_L\cos\theta_H}\right)^q \frac{n_{N+1}}{n_0}}{\left(-\frac{n_L\cos\theta_H}{n_H\cos\theta_L}\right)^q \frac{\cos\theta_{N+1}}{\cos\theta_0} + \left(-\frac{n_H\cos\theta_L}{n_L\cos\theta_H}\right)^q \frac{n_{N+1}}{n_0}} \quad (\lambda/4 \text{ stack, } p\text{-polarized}) \tag{4.68}$$

The quarter-wave multilayer considered in Example 4.6 can achieve extraordinarily high reflectivity. In the limit of $q \to \infty$, we have $t_p \to 0$ and $r_p \to -1$ (see Fig. 4.19), giving 100% reflection with a π phase shift.

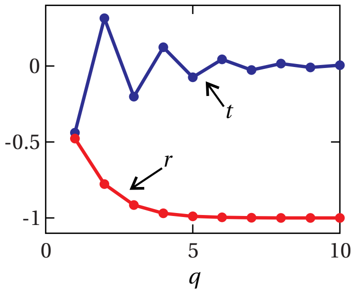


Figure 4.19 The transmission and reflection coefficients for a quarter wave stack as q is varied ($n_L = 1.38$ and $n_H = 2.32$).

Exercises

Exercises for 4.1 Double-Interface Problem Solved Using Fresnel Coefficients

P4.1 Use (4.4)–(4.7) to derive r_s^{tot} given in (4.12).

P4.2 Consider a 1μm thick coating of dielectric material ($n = 2$) on a piece of glass ($n = 1.5$). Use a computer to plot the magnitude of the overall Fresnel coefficient (4.11) from air into the glass at normal incidence. Plot as a function of incident wavelength in the range 200 nm to 800 nm, assuming the index remains constant over this range.

Exercises for 4.2 Transmittance through Double Interface at Subcritical Angles

P4.3 Verify that (4.14) simplifies to (4.15) assuming θ_1 and θ_2 are real.

P4.4 A light wave impinges at normal incidence on a thin glass plate with index n and thickness d.

(a) Show that the transmittance through the plate is

$$T^{\text{tot}} = \frac{1}{1 + \frac{\left(n^2-1\right)^2}{4n^2}\sin^2\left(\frac{2\pi n d}{\lambda_{\text{vac}}}\right)}$$

HINT: Find

$$r^{1\to 2} = r^{0\leftarrow 1} = -r^{0\to 1} = \frac{n-1}{n+1}$$

and then use

$$T^{0\to 1} = 1 - R^{0\to 1}$$

$$T^{1\to 2} = 1 - R^{1\to 2}$$

(b) If $n = 1.5$, what is the maximum and minimum possible transmittance through the plate?

(c) If the plate thickness is $d = 150\ \mu$m (same index as part (b)), what wavelengths transmit with maximum throughput? Express your answer as a formula involving an integer m.

P4.5 Show that the maximum reflectance possible from the front coating in Example 4.2 is 46%. Find the smallest possible d that accomplishes this for light with wavelength $\lambda_{\text{vac}} = 633$ nm.

Exercises for 4.3 Beyond Critical Angle: Tunneling of Evanescent Waves

P4.6 Re-compute (4.22) for the case of s-polarized light. Write the result in the same form as the final expression in (4.22).

Answer: $T_s^{\text{tot}} = \frac{1.44}{e^{4.44d/\lambda}+e^{-4.44d/\lambda}-0.560}$

L4.7 Consider s-polarized microwaves ($\lambda_{\text{vac}} = 3$ cm) encountering an air gap separating two paraffin wax prisms ($n = 1.5$). The 45° right-angle prisms are arranged with the geometry shown in Fig. 4.6. The presence of the second prism 'frustrates' the total internal reflection.

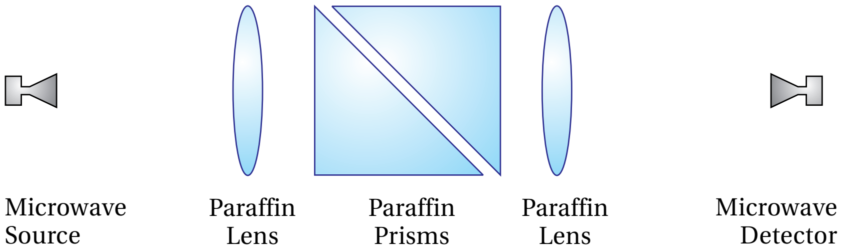


Figure 4.21

(a) Use a computer to plot the transmittance through the gap (i.e. the result of P4.6) as a function of separation d (normal to gap surface). Neglect reflections from other surfaces of the prisms.

(b) Measure the transmittance of the microwaves through the gap as a function of spacing d (normal to the surface) and superimpose the results on the graph of part (a). Figure 4.20 shows a plot of typical data taken with this setup. HINT: Ignore surface reflections by normalizing the measured power to a value of 1 when $d = 0$. (video)

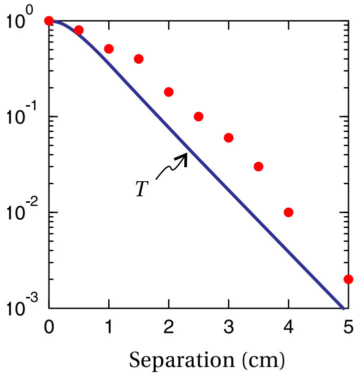


Figure 4.20 Theoretical vs. measured microwave transmission through wax prisms. Mismatch is presumably due to imperfections in microwave collimation and/or extraneous reflections.

Exercises for 4.6 Distinguishing Nearby Wavelengths in a Fabry-Perot Instrument

P4.8 A Fabry-Perot interferometer has silver-coated plates each with reflectance $R = 0.9$, transmittance $T = 0.05$, and absorbance $A = 0.05$. The plate separation is $d = 0.5$ cm with interior index $n_1 = 1$. Suppose that the wavelength being observed near normal incidence is 587 nm.

(a) What is the maximum and minimum transmittance through the interferometer?

(b) What are the free spectral range $\Delta\lambda_{\text{FSR}}$ and the fringe width $\Delta\lambda_{\text{FWHM}}$?

(c) What is the resolving power?

P4.9 Generate a plot like Fig. 4.13, showing the fringes you get in a Fabry-Perot etalon when θ_1 is varied. Let $T^{\max} = 1$, $F = 10$, $\lambda_{\text{vac}} = 633$ nm, $d = 1$ cm, and $n_1 = 1$.

(a) Plot T vs. θ_1 over the angular range used in Fig. 4.13.

(b) Suppose d is slightly different, say 1.00001 cm. Make a plot of $T^{\max}$ vs θ_1 for this situation.

P4.10 Consider the configuration depicted in Fig. 4.12, where the center of the diverging light beam $\lambda_{\text{vac}} = 633$ nm approaches the plates at normal incidence. Suppose that the spacing of the plates (near $d = 0.5$ cm) is just right to cause a bright fringe to occur at the center. Let $n_1 = 1$. Find the angle for the m^{th} circular bright fringe surrounding the central spot (the 0^{th} fringe corresponding to the center). HINT: $\cos\theta \cong 1 - \theta^2/2$. The answer has the form $a\sqrt{m}$; find the value of a.

L4.11 Characterize a Fabry-Perot etalon in the laboratory using a HeNe laser ($\lambda_{\text{vac}} = 633$ nm). Assume that the bandwidth $\Delta\lambda_{\text{HeNe}}$ of the HeNe laser is very narrow compared to the fringe width of the etalon $\Delta\lambda_{\text{FWHM}}$. Assume two identical reflective surfaces separated by 5.00 mm. Deduce the free spectral range $\Delta\lambda_{\text{FSR}}$, the fringe width $\Delta\lambda_{\text{FWHM}}$, the resolving power RP, and the reflecting finesse f. (video)

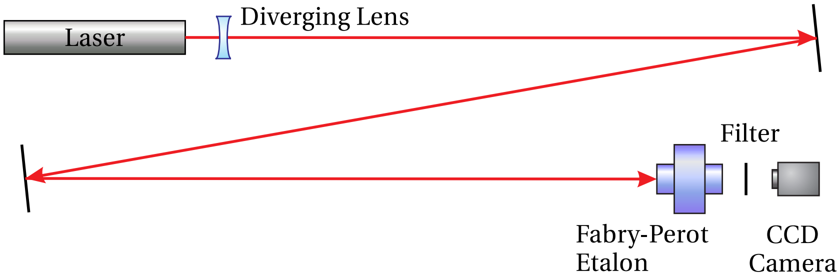


Figure 4.22

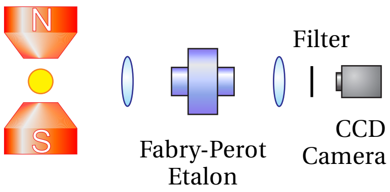


Figure 4.23

L4.12 Use the Fabry-Perot etalon characterized in the previous exercise to observe the Zeeman splitting of the yellow line $\lambda = 587.4$ nm emitted by a krypton lamp when a magnetic field is applied. As the line splits and moves through half of the free spectral range, the peak of the decreasing wavelength and the peak of the increasing wavelength meet on the screen. When this happens, by how much has each wavelength shifted? (video)

Exercises for 4.7 Multilayer Coatings

P4.13 (a) Write (4.42) through (4.47) for s-polarized light.

(b) From these equations, derive (4.61)–(4.63).

P4.14 Show that (4.64) for a single layer (i.e. two interfaces), is equivalent to (4.11). WARNING: This is more work than it may appear at first.

Exercises for 4.8 Periodic Multilayer Stacks

P4.15 (a) What should be the thickness of the high and the low index layers in a periodic high-reflector mirror? Let the light be p-polarized and strike the mirror surface at 45°. Take the indices of the layers be $n_{\mathrm{H}} = 2.32$ and $n_{\mathrm{L}} = 1.38$, deposited on a glass substrate with index $n = 1.5$. Let the wavelength be $\lambda_{\text{vac}} = 633$ nm.

(b) Find the reflectance R with 1, 2, 4, and 8 periods in the high-low stack.

P4.16 Find the high-reflector matrix for s-polarized light that corresponds to (4.66).

P4.17 Consider an anti-reflection coating designed for use at normal incidence between air ($n_0 = 1$) and glass ($n_g = 1.50$):

(a) Show that the reflectance of a single-layer $\lambda/4$ coating (where λ is the wavelength in n_1) is

$$R = \left(\frac{n_g - n_1^2}{n_g + n_1^2}\right)^2$$

(b) Show that for a two-coating arrangement (where n_1 and n_2 are each a $\lambda/4$ film), that

$$R = \left(\frac{n_2^2 - n_g n_1^2}{n_2^2 + n_g n_1^2}\right)^2$$

(c) Design anti-reflection coatings using the scheme in (a) and the scheme in (b). You have a choice of these common coating materials: ZnS ($n = 2.32$), CeF ($n = 1.63$) and MgF ($n = 1.38$). Find the recipe that gives you the lowest R in each case. (When considering scheme (b), be sure to specify which material is n_1 and which is n_2.)

P4.18 In this problem, we will see that the trick used in P4.17, employing a bilayer to improve anti reflection, doesn't get better with repeated bilayers. Consider a bilayer anti-reflection coating (each coating set for $\lambda/4$) using $n_1 = 1.38$ and $n_2 = 1.38$ applied to a glass substrate $n_g = 1.50$ at normal incidence. Suppose the coating thicknesses are optimized for $\lambda_{\text{vac}} = 550$ nm (in the middle of the visible range) and ignore possible variations of the indices with λ. Use a computer to plot $R(\lambda_{\text{air}})$ for 400 to 700 nm (visible range). Do this for a single bilayer (one layer of each coating), two bilayers, four bilayers, and 25 bilayers. HINT: You will see the good AR coating turn into a good HR coating.

Review, Chapters 1–4

True and False Questions

R1 T or F: The optical index of materials (not vacuum) varies with frequency.

R2 T or F: The frequency of light can change as it enters a different material (consider low intensity—no nonlinear effects).

R3 T or F: The entire expression $\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$ associated with a light field (both the real part and the imaginary parts) describes the physical wave.

R4 T or F: The *real* part of the refractive index cannot be less than one.

R5 T or F: s-polarized light and p-polarized light experience the *same* phase shift upon reflection from a material with complex index.

R6 T or F: When p-polarized light enters a material at Brewster's angle, the *intensity* of the transmitted beam is the same as the intensity of the incident beam.

R7 T or F: When light is incident upon a material interface at Brewster's angle, only one polarization can transmit.

R8 T or F: When light is incident upon a material interface at Brewster's angle, p-polarized light stimulates dipoles in the material to oscillate with orientation along $\mathbf{k}_r$.

R9 T or F: The critical angle for total internal reflection exists on both sides of a material interface.

R10 T or F: From a given location above a (smooth flat) surface of water, it is *possible* to see objects positioned anywhere under the water.

R11 T or F: From a given location beneath a (smooth flat) surface of water, it is *possible* to see objects positioned anywhere above the water.

R12 T or F: For incident angles beyond the critical angle for total internal reflection, the Fresnel coefficients t_s and t_p are both zero.

R13 T or F: Evanescent waves travel parallel to the surface interface on the transmitted side.

R14 T or F: For a given incident angle and value of n, there is only one single-layer coating thickness d that will minimize reflections.

R15 T or F: It is always possible *to completely eliminate* reflections using a single-layer antireflection coating if you are free to choose the coating thickness but not its index.

R16 T or F: When coating each surface of a lens with a single-layer antireflection coating (made of the same material), the thickness of the coating on the front of the lens will need to be different from the thickness of the coating on the back of the lens.

Problems

R17 (a) Write down Maxwell's equations from memory.

(b) Derive the wave equation for $\mathbf{E}$ under the assumptions that $\mathbf{J}_{\text{free}} = 0$ and $\mathbf{P} = \epsilon_0 \chi \mathbf{E}$ (which also implies $\nabla \cdot \mathbf{P} = 0$). Note: $\nabla \times (\nabla \times \mathbf{E}) = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E}$.

(c) Show by direct substitution that $\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)}$ is a solution to the wave equation. Find the resulting connection between k and ω. Give appropriate definitions for c and n, assuming that χ is real.

(d) If $\mathbf{k} = k\hat{\mathbf{z}}$ and $\mathbf{E}_0 = E_0 \hat{\mathbf{x}}$, find the associated $\mathbf{B}$-field.

(e) The Poynting vector is $\mathbf{S} = \mathbf{E} \times \mathbf{B}/\mu_0$. Derive an expression for $I \equiv \langle S \rangle_t$. HINT: You must use real fields.

R18 Consider an interface between two isotropic media where the incident field is defined by

$$\mathbf{E}_{\mathrm{i}} = \left[E_{\mathrm{i}}^{(p)} \left(\hat{\mathbf{y}} \cos\theta_{\mathrm{i}} - \hat{\mathbf{z}} \sin\theta_{\mathrm{i}} \right) + \hat{\mathbf{x}} E_{\mathrm{i}}^{(s)} \right] e^{i[k_{\mathrm{i}}(y \sin\theta_{\mathrm{i}} + z \cos\theta_{\mathrm{i}}) - \omega_{\mathrm{i}} t]}$$

The plane of incidence is shown in Fig. 4.24

(a) By inspection of the figure, write down similar expressions for the reflected and transmitted fields (i.e. $\mathbf{E}_{\mathrm{r}}$ and $\mathbf{E}_{\mathrm{t}}$).

(b) Find an expression relating $\mathbf{E}_{\mathrm{i}}$, $\mathbf{E}_{\mathrm{r}}$, and $\mathbf{E}_{\mathrm{t}}$ using the boundary condition at the interface. Also obtain the law of reflection and Snell's law.

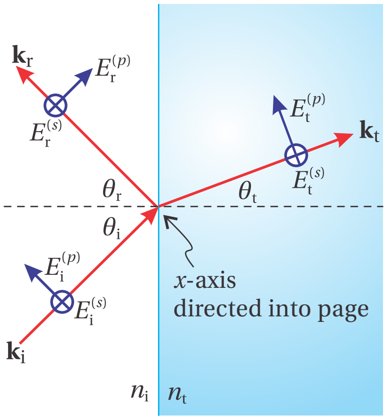


Figure 4.24

(c) The boundary condition requiring that the tangential component of **B** must be continuous leads to

$$n_\mathrm{i}(E_\mathrm{i}^{(p)} - E_\mathrm{r}^{(p)}) = n_t E_\mathrm{t}^{(p)}$$

$$n_\mathrm{i}(E_\mathrm{i}^{(s)} - E_\mathrm{r}^{(s)}) \cos\theta_\mathrm{i} = n_\mathrm{t} E_\mathrm{t}^{(s)} \cos\theta_\mathrm{t}$$

Use this and the results from part (b) to derive

$$r_p \equiv \frac{E_\mathrm{r}^{(p)}}{E_\mathrm{i}^{(p)}} = -\frac{\tan(\theta_\mathrm{i} - \theta_\mathrm{t})}{\tan(\theta_\mathrm{i} + \theta_\mathrm{t})}$$

R19 The Fresnel coefficients may be written

$$r_s \equiv \frac{E_\mathrm{r}^{(s)}}{E_\mathrm{i}^{(s)}} = \frac{\sin\theta_\mathrm{t}\cos\theta_\mathrm{i} - \sin\theta_\mathrm{i}\cos\theta_\mathrm{t}}{\sin\theta_\mathrm{t}\cos\theta_\mathrm{i} + \sin\theta_\mathrm{i}\cos\theta_\mathrm{t}}$$

$$t_s \equiv \frac{E_\mathrm{t}^{(s)}}{E_\mathrm{i}^{(s)}} = \frac{2\sin\theta_\mathrm{t}\cos\theta_\mathrm{i}}{\sin\theta_\mathrm{t}\cos\theta_\mathrm{i} + \sin\theta_\mathrm{i}\cos\theta_\mathrm{t}}$$

$$r_p \equiv \frac{E_\mathrm{r}^{(p)}}{E_\mathrm{i}^{(p)}} = \frac{\cos\theta_\mathrm{t}\sin\theta_\mathrm{t} - \cos\theta_\mathrm{i}\sin\theta_\mathrm{i}}{\cos\theta_\mathrm{t}\sin\theta_\mathrm{t} + \cos\theta_\mathrm{i}\sin\theta_\mathrm{i}}$$

$$t_p \equiv \frac{E_\mathrm{t}^{(p)}}{E_\mathrm{i}^{(p)}} = \frac{2\cos\theta_\mathrm{i}\sin\theta_\mathrm{t}}{\cos\theta_\mathrm{t}\sin\theta_\mathrm{t} + \cos\theta_\mathrm{i}\sin\theta_\mathrm{i}}$$

(a) Make substitutions from Snell's law to show what each of these equations reduces to when $\theta_\mathrm{i} = 0$. Express you answers in terms of n_i and n_t.

(b) What percent of light (intensity) reflects from a glass surface ($n = 1.5$) when light enters from air ($n = 1$) at normal incidence?

(c) What percent of light reflects from the glass surface when light exits into air at normal incidence?

R20 Light goes through a glass prism with optical index $n = 1.55$. The light enters at Brewster's angle and exits at normal incidence as shown in Fig. 4.25.

(a) *Derive* and calculate Brewster's angle θ_B. You may use the results of R18 (c).

(b) Calculate ϕ.

(c) What percent of the light (power) goes all the way through the prism if it is p-polarized? You may use the expression employed in R19(c).

(d) Repeat part (c) for s-polarized light.

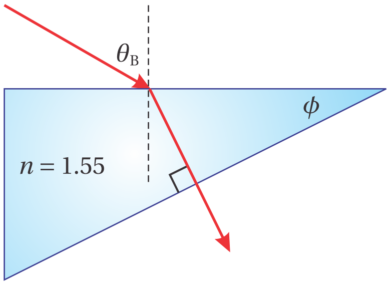


Figure 4.25

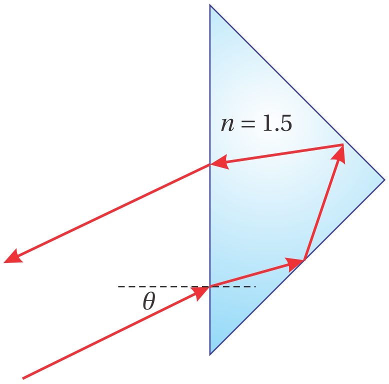


Figure 4.26

R21 A 45°- 90°- 45° prism is a good device for reflecting a beam of light parallel to the initial beam (see Fig. 4.26). The exiting beam will be parallel to the entering beam even when the incoming beam is not normal to the front surface (although it needs to be in the plane of the drawing).

(a) How large an angle θ can be tolerated before there is no longer total internal reflection at both interior surfaces? Assume $n = 1$ outside of the prism and $n = 1.5$ inside.

(b) If the light enters and leaves the prism at normal incidence, what will the difference in phase be between the s and p-polarizations? You may use the Fresnel coefficients provided in R19.

R22 A thin glass plate with index $n = 1.5$ is oriented at Brewster's angle so that p-polarized light with wavelength $\lambda_{\text{vac}} = 500$ nm goes through with 100% transmittance.

(a) What is the minimum thickness that will make the reflection of s-polarized light be maximum?

(b) What is the total transmittance T_s^{tot} for this thickness assuming s-polarized light?

R23 Consider an ideal Fabry-Perot interferometer with

$$T^{\text{tot}} = \frac{T^{\text{max}}}{1 + F\sin^2\left(\frac{\Phi}{2}\right)} \quad , \quad T^{\text{max}} = \frac{T^2}{(1-R)^2} \quad , \quad F = \frac{4R}{(1-R)^2}$$

$$\text{and} \quad \Phi = \frac{4\pi n_1 d}{\lambda_{\text{vac}}}\cos\theta_1 + 2\phi_r$$

(a) Derive the free spectral range

$$\Delta\lambda_{\text{FSR}} = \frac{\lambda_{\text{vac}}^2}{2nd\cos\theta_1}$$

(b) Derive the fringe width

$$\Delta\lambda_{\text{FWHM}} = \frac{\lambda_{\text{vac}}^2}{\pi\sqrt{F}n_1 d\cos\theta_1}$$

(c) Give the reflecting finesse $f = \Delta\lambda_{\text{FSR}}/\Delta\lambda_{\text{FWHM}}$.

R24 For a Fabry-Perot etalon, let $R = 0.90$, $\lambda_{\text{vac}} = 500$ nm, $n = 1$, and $d = 5.0$ mm.

(a) Suppose that a maximum transmittance occurs at the angle $\theta = 0$. What is the nearest angle where the transmittance will be half of the maximum transmittance? You may assume that $\cos\theta \cong 1 - \theta^2/2$.

(b) You desire to use a Fabry-Perot etalon to view the light from a large diffuse source rather than a point source. Draw a diagram depicting where lenses should be placed, indicating relevant distances. Explain briefly how it works.

R25 The p-polarized matrix equation relating reflected and transmitted fields to the incident field impinging on multilayer interface is

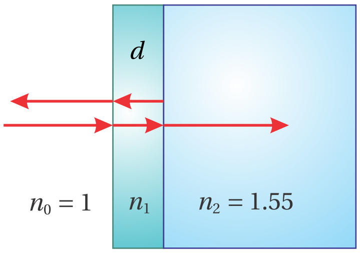


Figure 4.27

$$\begin{bmatrix} 1 \\ E_{0\leftarrow}^{(p)}/E_{0\rightarrow}^{(p)} \end{bmatrix} = A^{(p)} \begin{bmatrix} E_{N+1\rightarrow}^{(p)}/E_{0\rightarrow}^{(p)} \\ 0 \end{bmatrix}$$

where

$$A^{(p)} = \frac{1}{2n_0\cos\theta_0}\begin{bmatrix} n_0 & \cos\theta_0 \\ n_0 & -\cos\theta_0 \end{bmatrix}\left(\prod_{j=1}^{N} M_j^{(p)}\right)\begin{bmatrix} \cos\theta_{N+1} & 0 \\ n_{N+1} & 0 \end{bmatrix}$$

$$M_j^{(p)} = \begin{bmatrix} \cos\beta_j & -i\sin\beta_j\cos\theta_j/n_j \\ -in_j\sin\beta_j/\cos\theta_j & \cos\beta_j \end{bmatrix} \qquad \beta_j = k_j d_j\cos\theta_j$$

(a) If the layer is an antireflective coating applied between air ($n_0 = 1$) and glass ($n_2 = 1.55$) designed to work at normal incidence. What is the minimum thickness the coating should have? HINT: It is less work if you can figure this out without referring to the above equation.

(b) If there is just one layer of material, show that at normal incidence the above matrix equation for the thickness chosen in (a) reduces to

$$\begin{bmatrix} 1 \\ E_{0\leftarrow}^{(p)}/E_{0\rightarrow}^{(p)} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} -i\left(\frac{n_1}{n_0}+\frac{n_2}{n_1}\right) & 0 \\ i\left(\frac{n_1}{n_0}-\frac{n_2}{n_1}\right) & 0 \end{bmatrix}\begin{bmatrix} E_{2\rightarrow}^{(p)}/E_{0\rightarrow}^{(p)} \\ 0 \end{bmatrix}$$

(c) Assuming the parameters in part (b), find the index of refraction n_1 that will make the reflectance be zero.

Selected Answers

R19: (b) 4% (c) 4%.

R20: (b) 33°, (c) 95%, (d) 79%.

R21: (a) 4.8°, (b) 74°.

R22: (a) 100 nm. (b) 0.55.

R24: (a) 0.074°.

P25: (c) 1.24.

Chapter 5

Propagation in Anisotropic Media

To this point, we have considered only *isotropic* media where the susceptibility $\chi(\omega)$ (and hence the index of refraction) is the same for all propagation directions and polarizations. In *anisotropic* materials, such as crystals, it is possible for light to experience a different index of refraction depending on the alignment of the electric field **E** (i.e. polarization). This difference in the index of refraction occurs when the direction and strength of the induced dipoles depends on the lattice structure of the material in addition to the propagating field.[1] The unique properties of anisotropic materials make them important elements in many optical systems.

In section 5.1 we discuss how to connect **E** and **P** in anisotropic media using a *susceptibility tensor*. In section 5.2 we apply Maxwell's equations to a plane wave traveling in a crystal. The analysis leads to *Fresnel's equation*, which relates the components of the **k**-vector to the components of the susceptibility tensor. In section 5.3 we apply Fresnel's equation to a *uniaxial* crystal (e.g. quartz, sapphire) where $\chi_x = \chi_y \neq \chi_z$. In the context of a uniaxial crystal, we show that the Poynting vector and the **k**-vector are generally not parallel.

More than a century before Fresnel, Christian Huygens successfully described *birefringence* in crystals using the idea of elliptical wavelets. His method gives the direction of the Poynting vector associated with the *extraordinary* ray in a crystal. It was Huygens who coined the term 'extraordinary' since one of the rays in a *birefringent* material appeared not to obey Snell's law. Actually, the **k**-vector always obeys Snell's law, but in a crystal, the **k**-vector points in a different direction than the Poynting vector, which delivers the energy seen by an observer. Huygens' approach is outlined in Appendix 5.D.

5.1 Constitutive Relation in Crystals

In an anisotropic crystal, asymmetries in the lattice can cause the medium polarization **P** to respond in a different direction than the electric field **E** (i.e. $\mathbf{P} \neq \epsilon_0 \chi \mathbf{E}$).

---

[1] Not all crystals are anisotropic. For instance, crystals with a cubic lattice structure (such as NaCl) are highly symmetric and respond to electric fields the same in any direction.

However, at low intensities the response of materials is still linear (or proportional) to the strength of the electric field. The linear *constitutive relation* which connects **P** to **E** in a crystal can be expressed in its most general form as

$$\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} = \epsilon_0 \begin{bmatrix} \chi_{xx} & \chi_{xy} & \chi_{xz} \\ \chi_{yx} & \chi_{yy} & \chi_{yz} \\ \chi_{zx} & \chi_{zy} & \chi_{zz} \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} \tag{5.1}$$

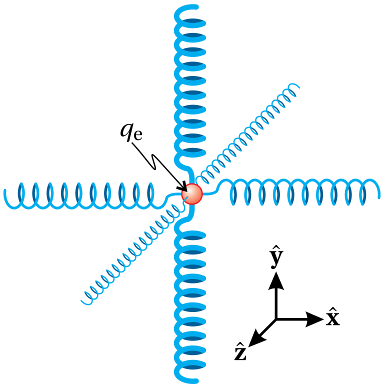


Figure 5.1 A physical model of an electron bound in a crystal lattice with the coordinate system specially chosen along the principal axes so that the susceptibility tensor takes on a simple form.

The matrix in (5.1) is called the *susceptibility tensor*. To visualize the behavior of electrons in such a material, we imagine each electron bound as though by tiny springs with different strengths in different dimensions to represent the anisotropy (see Fig. 5.1). When an external electric field is applied, the electron experiences a force that moves it from its equilibrium position. The 'springs' (actually the electric force from ions bound in the crystal lattice) exert a restoring force, but the restoring force is not equal in all directions—the electron tends to move more along the dimension of the weaker spring. The displaced electron creates a microscopic dipole, but the asymmetric restoring force causes **P** to be in a direction different than **E** as depicted in Fig. 5.2.

To understand the geometrical interpretation of the many coefficients χ_{ij}, assume, for example, that the electric field is directed along the x-axis (i.e. $E_y = E_z = 0$) as depicted in Fig. 5.2. In this case, the three equations encapsulated in (5.1) reduce to

$$\begin{aligned} P_x &= \epsilon_0 \chi_{xx} E_x \\ P_y &= \epsilon_0 \chi_{yx} E_x \\ P_z &= \epsilon_0 \chi_{zx} E_x \end{aligned}$$

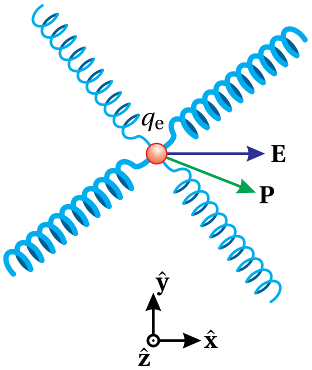


Figure 5.2 The applied field **E** and the induced polarization **P** in general are not parallel in a crystal lattice.

Notice that the coefficient χ_{xx} connects the strength of **P** in the $\hat{\mathbf{x}}$ direction with the strength of **E** in that same direction, just as in the isotropic case. The other two coefficients (χ_{yx} and χ_{zx}) describe the amount of polarization **P** produced in the $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ directions by the electric field component in the x-dimension. Likewise, the other coefficients with mixed subscripts in (5.1) describe the contribution to **P** in one dimension made by an electric field component in another dimension.

As you might imagine, working with nine susceptibility coefficients can get complicated. Fortunately, we can greatly reduce the complexity of the description by a judicious choice of coordinate system. In Appendix 5.A we explain how conservation of energy requires that the susceptibility tensor (5.1) for typical nonabsorbing crystals be real and symmetric (i.e. $\chi_{ij} = \chi_{ji}$).[2]

Appendix 5.B shows that, given a real symmetric tensor, it is always possible to choose a coordinate system for which off-diagonal elements vanish. This is true even if the lattice planes in the crystal are not mutually orthogonal (e.g. rhombus, hexagonal, etc.). We will imagine that this rotation of coordinates has been accomplished. In other words, we can let the crystal itself dictate the orientation of the coordinate system, aligned to the *principal axes* of the crystal for which the off-diagonal elements of (5.1) are zero

---

[2] By 'typical' we mean that the crystal does not exhibit optical activity. Optically active crystals have a complex susceptibility tensor, even when no absorption takes place. Conservation of energy in this more general case requires that the susceptibility tensor be Hermitian ($\chi_{ij} = \chi^*_{ji}$).

With the coordinate system aligned to the principal axes, the constitutive relation for a nonabsorbing crystal simplifies to

$$\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} = \epsilon_0 \begin{bmatrix} \chi_x & 0 & 0 \\ 0 & \chi_y & 0 \\ 0 & 0 & \chi_z \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} \tag{5.2}$$

or without the matrix notation (since it no longer offers much convenience)

$$\mathbf{P} = \hat{\mathbf{x}}\epsilon_0\chi_x E_x + \hat{\mathbf{y}}\epsilon_0\chi_y E_y + \hat{\mathbf{z}}\epsilon_0\chi_z E_z \tag{5.3}$$

By assumption, χ_x, χ_y, and χ_z are all real. (We have dropped the double subscript; χ_x stands for χ_{xx}, etc.)

5.2 Plane Wave Propagation in Crystals

We consider a plane wave with frequency ω propagating in a crystal. In a manner similar to our previous analysis of plane waves propagating in isotropic materials, we write as trial solutions

$$\begin{aligned} \mathbf{E} &= \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \\ \mathbf{B} &= \mathbf{B}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \\ \mathbf{P} &= \mathbf{P}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \end{aligned} \tag{5.4}$$

where restrictions on $\mathbf{E}_0$, $\mathbf{B}_0$, $\mathbf{P}_0$, and $\mathbf{k}$ are yet to be determined. As usual, the phase of each wave is included in the amplitudes $\mathbf{E}_0$, $\mathbf{B}_0$, and $\mathbf{P}_0$, whereas $\mathbf{k}$ is real in accordance with our assumption of no absorption.

We can make a quick observation about the behavior of these fields by applying Maxwell's equations directly. Gauss's law for electric fields requires

$$\nabla\cdot(\epsilon_0\mathbf{E}+\mathbf{P}) = \mathbf{k}\cdot(\epsilon_0\mathbf{E}+\mathbf{P}) = 0 \tag{5.5}$$

and Gauss's law for magnetism gives

$$\nabla\cdot\mathbf{B} = \mathbf{k}\cdot\mathbf{B} = 0 \tag{5.6}$$

We immediately notice the following peculiarity: From its definition, the Poynting vector $\mathbf{S} \equiv \mathbf{E}\times\mathbf{B}/\mu_0$ is perpendicular to both $\mathbf{E}$ and $\mathbf{B}$, and by (5.6) the $\mathbf{k}$-vector is perpendicular to $\mathbf{B}$. However, by (5.5) the $\mathbf{k}$-vector is not necessarily perpendicular to $\mathbf{E}$, since in general $\mathbf{k}\cdot\mathbf{E} \neq 0$ if $\mathbf{P}$ points in a direction other than $\mathbf{E}$. Therefore, $\mathbf{k}$ and $\mathbf{S}$ are not necessarily parallel in a crystal. In other words, the flow of energy and the direction of the phase-front propagation can be different in anisotropic media.

Our main goal here is to relate the **k**-vector to the susceptibility parameters χ_x, χ_y, and χ_z. To do this, we plug our trial plane-wave fields into the wave equation (1.40). Under the assumption $\mathbf{J}_{\text{free}} = 0$, we have

$$\nabla^2\mathbf{E} - \mu_0\epsilon_0\frac{\partial^2\mathbf{E}}{\partial t^2} = \mu_0\frac{\partial^2\mathbf{P}}{\partial t^2} + \nabla(\nabla\cdot\mathbf{E}) \tag{5.7}$$

Derivation of the dispersion relation in crystals

We begin by substituting the trial solutions (5.4) into the wave equation (5.7). After carrying out the derivatives we find

$$k^2\mathbf{E} - \omega^2\mu_0(\epsilon_0\mathbf{E} + \mathbf{P}) = \mathbf{k}(\mathbf{k}\cdot\mathbf{E}) \tag{5.8}$$

Inserting the constitutive relation (5.3) for crystals into (5.8) yields

$$k^2\mathbf{E} - \omega^2\mu_0\epsilon_0\left[\left(1+\chi_x\right)E_x\hat{\mathbf{x}} + \left(1+\chi_y\right)E_y\hat{\mathbf{y}} + \left(1+\chi_z\right)E_z\hat{\mathbf{z}}\right] = \mathbf{k}(\mathbf{k}\cdot\mathbf{E}) \tag{5.9}$$

This relationship is unwieldy because of the mix of electric field components that appear in the expression. This was not a problem when we investigated isotropic materials for which the **k**-vector is perpendicular to **E**, making the right-hand side of the equations zero. However, there is a trick for dealing with this.

Relation (5.9) actually contains three equations, one for each dimension. Explicitly, these equations are

$$\left[k^2 - \frac{\omega^2}{c^2}\left(1+\chi_x\right)\right]E_x = k_x(\mathbf{k}\cdot\mathbf{E}) \tag{5.10}$$

$$\left[k^2 - \frac{\omega^2}{c^2}\left(1+\chi_y\right)\right]E_y = k_y(\mathbf{k}\cdot\mathbf{E}) \tag{5.11}$$

and

$$\left[k^2 - \frac{\omega^2}{c^2}\left(1+\chi_z\right)\right]E_z = k_z(\mathbf{k}\cdot\mathbf{E}) \tag{5.12}$$

We have replaced the constants $\mu_0\epsilon_0$ with $1/c^2$ in accordance with (1.42). We multiply (5.10)–(5.12) respectively by k_x, k_y, and k_z and also move the factor in square brackets in each equation to the denominator on the right-hand side. Then by adding the three equations together we get

$$\frac{k_x^2(\mathbf{k}\cdot\mathbf{E})}{\left[k^2 - \frac{\omega^2(1+\chi_x)}{c^2}\right]} + \frac{k_y^2(\mathbf{k}\cdot\mathbf{E})}{\left[k^2 - \frac{\omega^2(1+\chi_y)}{c^2}\right]} + \frac{k_z^2(\mathbf{k}\cdot\mathbf{E})}{\left[k^2 - \frac{\omega^2(1+\chi_z)}{c^2}\right]} = k_xE_x + k_yE_y + k_zE_z = (\mathbf{k}\cdot\mathbf{E}) \tag{5.13}$$

Now $\mathbf{k}\cdot\mathbf{E}$ appears in every term and can be divided away. This gives the dispersion relation (unencumbered by field components):

$$\frac{k_x^2}{\left[k^2c^2/\omega^2 - \left(1+\chi_x\right)\right]} + \frac{k_y^2}{\left[k^2c^2/\omega^2 - \left(1+\chi_y\right)\right]} + \frac{k_z^2}{\left[k^2c^2/\omega^2 - \left(1+\chi_z\right)\right]} = \frac{\omega^2}{c^2} \tag{5.14}$$

As a final touch, we have multiplied the equation through by ω^2/c^2

The dispersion relation (5.14) allows us to find a suitable **k**, given values for ω, χ_x, χ_y, and χ_z. Actually, it only restricts the magnitude of **k**; we must still decide on a direction for the wave to travel (i.e. we must choose the ratios between k_x, k_y, and k_z). To remind ourselves of this fact, we introduce a unit vector that points in the direction of **k**

$$\mathbf{k} = k_x\hat{\mathbf{x}} + k_y\hat{\mathbf{y}} + k_z\hat{\mathbf{z}} = k\left(u_x\hat{\mathbf{x}} + u_y\hat{\mathbf{y}} + u_z\hat{\mathbf{z}}\right) = k\hat{\mathbf{u}} \tag{5.15}$$

With this unit vector inserted, the dispersion relation (5.14) for plane waves in a crystal becomes

$$\frac{u_x^2}{\left[k^2c^2/\omega^2 - \left(1+\chi_x\right)\right]} + \frac{u_y^2}{\left[k^2c^2/\omega^2 - \left(1+\chi_y\right)\right]} + \frac{u_z^2}{\left[k^2c^2/\omega^2 - \left(1+\chi_z\right)\right]} = \frac{\omega^2}{k^2c^2} \tag{5.16}$$

We may define refractive index as the ratio of the speed of light in vacuum c to the speed of phase propagation in a material ω/k (see P1.9). The relation introduced for isotropic media (i.e. (2.19) for real index) remains appropriate. That is

$$n = \frac{kc}{\omega} \tag{5.17}$$

This familiar relationship between k and ω, in the case of a crystal, depends on the direction of propagation in accordance with (5.16).

Inspired by (2.30), we will find it helpful to introduce several refractive-index parameters:

$$\begin{aligned} n_x &\equiv \sqrt{1+\chi_x} \\ n_y &\equiv \sqrt{1+\chi_y} \\ n_z &\equiv \sqrt{1+\chi_z} \end{aligned} \tag{5.18}$$

With these definitions (5.17)-(5.18), the dispersion relation (5.16) becomes

$$\frac{u_x^2}{\left(n^2-n_x^2\right)} + \frac{u_y^2}{\left(n^2-n_y^2\right)} + \frac{u_z^2}{\left(n^2-n_z^2\right)} = \frac{1}{n^2} \tag{5.19}$$

This is called *Fresnel's equation*[3] (not to be confused with the Fresnel coefficients studied in chapter 3). The relationship contains the yet unknown index n that varies with the direction of the **k**-vector (i.e. the direction of the unit vector $\hat{\mathbf{u}}$).

After multiplying through by all of the denominators (and after a fortuitous cancelation owing to $u_x^2 + u_y^2 + u_z^2 = 1$), Fresnel's equation (5.19) can be rewritten as a quadratic in n^2. The two solutions are

$$n^2 = \frac{B \pm \sqrt{B^2 - 4AC}}{2A} \tag{5.20}$$

---

[3] To better distinguish from the Fresnel coefficients, sometimes this is called *Fresnel's equation of wave normals*. See *Principles of Optics, 7th Ed.*, Born and Wolf, p. 796.

where

$$A \equiv u_x^2 n_x^2 + u_y^2 n_y^2 + u_z^2 n_z^2 \tag{5.21}$$

$$B \equiv u_x^2 n_x^2 \left(n_y^2 + n_z^2\right) + u_y^2 n_y^2 \left(n_x^2 + n_z^2\right) + u_z^2 n_z^2 \left(n_x^2 + n_y^2\right) \tag{5.22}$$

$$C \equiv n_x^2 n_y^2 n_z^2 \tag{5.23}$$

The upper and lower signs (+ and −) in (5.20) give two positive solutions for n^2. The positive square root of these solutions yields two physical values for n. It turns out that each of the two values for n is associated with a polarization direction of the electric field, given a propagation direction **k**. A broader analysis carried out in appendix 5.C renders the orientation of the electric fields, whereas here we only show how to find the two values of n. We refer to the two indices as the slow and fast index, since the waves associated with each propagate at speed $v = c/n$.

In the special cases of propagation along one of the principal axes of the crystal, the index n takes on two of the values n_x, n_y, or n_z, depending on which are orthogonal to the direction of propagation.

Example 5.1

Calculate the two possible values for the index of refraction when **k** is in the $\hat{\mathbf{z}}$ direction (in the crystal principal frame).

Solution: With $u_z = 1$ and $u_x = u_y = 0$ we have

$$A = n_z^2; \quad B = n_z^2 \left(n_x^2 + n_y^2\right); \quad C = n_x^2 n_y^2 n_z^2$$

The square-root term is then

$$\begin{aligned}\sqrt{B^2 - 4AC} &= \sqrt{n_z^4 \left(n_x^4 + 2n_x^2 n_y^2 + n_y^4\right) - 4n_x^2 n_y^2 n_z^4} \\ &= \sqrt{n_z^4 \left(n_x^2 - n_y^2\right)^2} \\ &= n_z^2 \left(n_x^2 - n_y^2\right)\end{aligned}$$

Inserting this expression into (5.20), we find the two values for the index:

$$n = n_x, n_y$$

The index n_x is experienced by light whose electric field points in the x-dimension, and the index n_y is experienced by light whose electric field points in the y-dimension (see appendix 5.C).

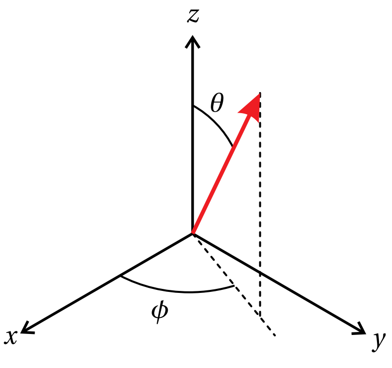


Figure 5.3 Spherical coordinates.

Before moving on, let us briefly summarize what has been accomplished so far. Given values for χ_x, χ_y, and χ_z associated with light in a crystal at a given frequency, you can define the indices n_x, n_y, and n_z, according to (5.18). Next, a direction for the **k**-vector is chosen (i.e. u_x, u_y, and u_z). This direction generally has two values for the index of refraction associated with it, found using Fresnel's equation (5.20). Each index is associated with a specific polarization direction for the electric field as outlined in appendix 5.C. Every propagation direction $\hat{\mathbf{u}}$ has its own natural set of polarization components for the electric field. The two polarization components travel at different speeds, even though the frequency is the same. This is known as *birefringence*.

5.3 Biaxial and Uniaxial Crystals

All anisotropic crystals have certain special propagation directions where the two values for n from Fresnel's equation are equal. These directions are referred to as the *optic axes* of the crystal. The optic axes do not necessarily coincide with the principal axes $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$. When propagation is along an optic axis, all polarization components experience the same index of refraction. If the values of n_x, n_y, and n_z are all unique, a crystal will have two optic axes, and hence is referred to as a *biaxial* crystal.

It is often convenient to use spherical coordinates to represent the components of $\hat{\mathbf{u}}$ (see to Fig. 5.3):

$$\begin{aligned} u_x &= \sin\theta\cos\phi \\ u_y &= \sin\theta\sin\phi \\ u_z &= \cos\theta \end{aligned} \tag{5.24}$$

Here θ is the polar angle measured from the z-axis of the crystal and ϕ is the azimuthal angle measured from the x-axis of the crystal. These equations emphasize the fact that there are only two degrees of freedom when specifying propagation direction (θ and ϕ). It is important to remember that these angles must be specified in the frame of the crystal's principal axes, which are often not aligned with the faces of a cut crystal in an optical setup.

By convention, we order the principal axes for biaxial crystals so that $n_x < n_y < n_z$. Under this convention, the two optic axes occur in the x-z plane ($\phi = 0$) at two values of the polar angle θ, measured from the z-axis (see P5.4):

$$\cos\theta = \pm\frac{n_x}{n_y}\sqrt{\frac{n_z^2 - n_y^2}{n_z^2 - n_x^2}} \qquad \text{(Optic axes directions, biaxial crystal)} \tag{5.25}$$

The index of refraction for light traveling along either optic axis is n_y. This results from the following two facts: the optic axis is in the x-z plane and light traveling in this plane can be polarized in the $\hat{\mathbf{y}}$-direction; and all polarization components for light traveling along the optic axis have the same index of refraction.

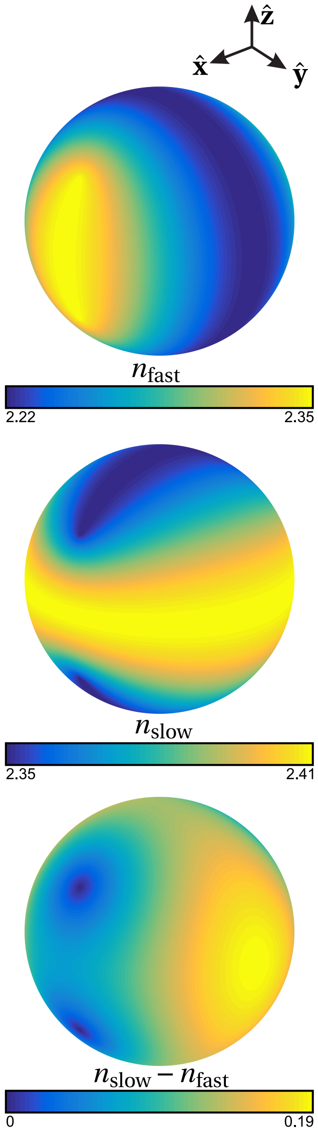


Figure 5.4 The fast and slow refractive indices (and their difference) as a function of direction for potassium niobate ($KNbO_3$) at $\lambda = 500$ nm ($n_x = 2.22$, $n_y = 2.35$, and $n_z = 2.41$) .

For arbitrary propagation directions the two indices of refraction are found using Fresnel's equation (5.20). The smaller value is commonly referred to as the 'fast' index and the larger value the 'slow' index. Figure 5.4 shows the two refractive indices (i.e. the solutions to Fresnel's equation) for a biaxial crystal plotted with color shading on the surface of a sphere. Each point on the sphere represents a different θ and ϕ. The two optic axes are apparent in the plot of the difference between n_{slow} and n_{fast}; the two indices coincide when propagation is along either optic axis.

For the remainder of this chapter, we will focus on the simpler case of *uniaxial* crystals. In uniaxial crystals two of the susceptibility coefficients χ_x, χ_y, and χ_z are the same. In this case, there is only one optic axis for the crystal. By convention, in uniaxial crystals we label the dimension that has the unique susceptibility as the z-axis (i.e. $\chi_x = \chi_y \neq \chi_z$). This makes the z-axis the optic axis. The unique index of refraction is called the *extraordinary index*

$$n_z = n_{\text{e}} \tag{5.26}$$

and the other index is called the *ordinary* index

$$n_x = n_y = n_{\text{o}} \tag{5.27}$$

These names were coined by Huygens, one of the early scientists to study light in crystals (see appendix 5.D). A uniaxial crystal with $n_{\text{e}} > n_{\text{o}}$ is referred to as a *positive crystal*, and one with $n_{\text{e}} < n_{\text{o}}$ is referred to as a *negative crystal.*

To calculate the index of refraction for a wave propagating in a uniaxial crystal, we use definitions (5.26) and (5.27) along with the spherical representation of $\hat{\mathbf{u}}$ (5.24) in Fresnel's equation (5.20) to find the following two values for n (see P5.5):

$$n = n_{\text{o}} \qquad \text{(uniaxial crystal)} \tag{5.28}$$

and

$$n = n_{\text{e}}(\theta) \equiv \frac{n_{\text{o}} n_{\text{e}}}{\sqrt{n_{\text{o}}^2 \sin^2\theta + n_{\text{e}}^2 \cos^2\theta}} \qquad \text{(uniaxial crystal)} \tag{5.29}$$

The angularly-dependent index $n_{\text{e}}(\theta)$ in (5.29) is also commonly referred to as the extraordinary index, the same name used for the constant n_{e}. (Note that $n_{\text{e}} = n_z = n_{\text{e}}(\theta = 0)$.) While this nomenclature has the potential for confusion, the practice is so common that we will perpetuate it here. We will write $n_{\text{e}}(\theta)$ when the angle dependent quantity specified by (5.29) is required, and write n_{e} in formulas where the constant (5.26) is called for (as in the right-hand side of (5.29)). Notice that $n_{\text{e}}(\theta)$ depends only on θ (the polar angle measured from the optic axis $\hat{\mathbf{z}}$) and not ϕ (the azimuthal angle). Figure 5.5 shows the two refractive indices (5.28) and (5.29) as a function θ and ϕ. Since $n_{\text{e}}(\theta)$ has no ϕ dependence and n_{o} is constant, the variation with direction is much simpler than for the biaxial case.

As outlined in appendix 5.C, the index n_{o} corresponds to an electric field component that points perpendicular to the plane containing $\hat{\mathbf{u}}$ and $\hat{\mathbf{z}}$. For example, if $\hat{\mathbf{u}}$ is in the x-z plane, n_{o} is associated with light polarized in the y-dimension. On the other hand, the index $n_{\text{e}}(\theta)$ corresponds to field polarization that lies within the plane containing $\hat{\mathbf{u}}$ and $\hat{\mathbf{z}}$. In this case, the polarization component is directed partially along the optic axis (i.e. it has a z-component). That is why (5.29) gives for the refractive index a mixture of n_{o} and n_{e}. If $\theta = 0$, then the $\mathbf{k}$-vector is directed exactly along the optic axis, and $n_{\text{e}}(\theta)$ reduces to n_{o} so that both polarization components experience same index n_{o}.

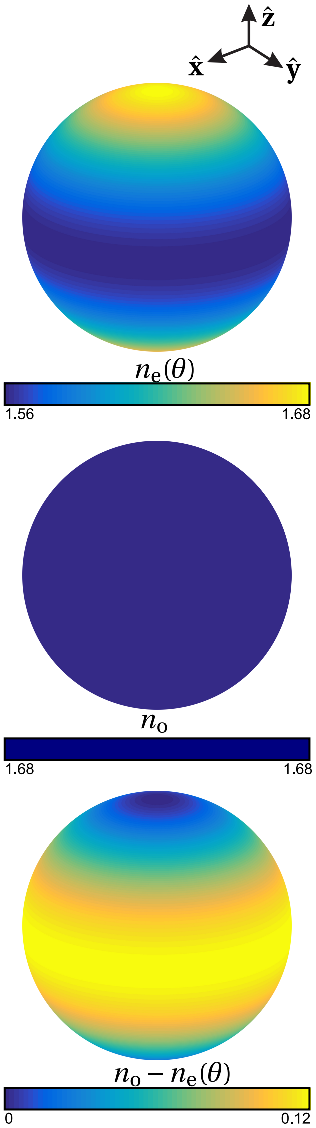


Figure 5.5 The extraordinary and ordinary refractive indices (and their difference) as a function of direction for beta barium borate (BBO) at $\lambda = 500$ nm ($n_{\text{o}} = 1.68$ and $n_{\text{e}} = 1.56$).

5.4 Refraction at a Uniaxial Crystal Surface

Next we consider refraction as light enters a uniaxial crystal. Snell's law (3.7) describes the connection between the **k**-vectors incident upon and transmitted through the surface. We must consider separately the portion of the light that experiences the ordinary index and the portion that experiences the extraordinary index. Because of the different indices, the ordinary and extraordinary polarized light refract into the crystal at two different angles; they travel at two different velocities in the crystal; and they have two different wavelengths in the crystal.

If we assume that the index outside of the crystal is one, Snell's law for the ordinary polarization is

$$\sin\theta_{\mathrm{i}} = n_{\mathrm{o}} \sin\theta_{\mathrm{t}} \qquad \text{(ordinary polarized light)} \qquad (5.30)$$

where n_{o} is the ordinary index inside the crystal. The extraordinary polarized light also obeys Snell's law, but now the index of refraction in the crystal depends on direction of propagation inside the crystal *relative to the optic axis*. Snell's law for the extraordinary polarization is

$$\sin\theta_{\mathrm{i}} = n_{\mathrm{e}}(\theta') \sin\theta_t \qquad \text{(extraordinary polarized light)} \qquad (5.31)$$

where θ' is the angle between the optic axis inside the crystal and the direction of propagation in the crystal (given by θ_{t} in the plane of incidence). When the optic axis is at an arbitrary angle with respect to the surface the relationship between θ' and θ_{t} is cumbersome. We will examine Snell's law only for the specific case when the optic axis is perpendicular to the crystal surface, for which $\theta_{\mathrm{t}} = \theta'$.

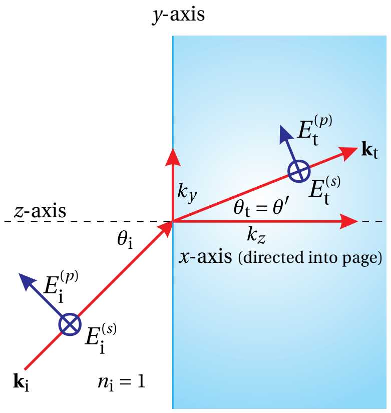


Figure 5.6 Propagation of light in a uniaxial crystal with its optic axis perpendicular to the surface.

Example 5.2

Examine Snell's law for a uniaxial crystal with optic axis perpendicular to the surface.

Solution: Refer to Fig. 5.6. With the optic axis perpendicular to the surface, if the light hits the crystal surface straight on, the index of refraction is n_{o}, regardless of the orientation of polarization since $\theta' = \theta_{\mathrm{t}} = 0$. When the light strikes the surface at an angle, s-polarized light continues to experience the index n_{o}, while p-polarized light experiences the extraordinary index $n_{\mathrm{e}}(\theta_{\mathrm{t}})$. [4]

When we insert (5.29) into Snell's law (5.31) with $\theta' = \theta_{\mathrm{t}}$, the expression can be inverted to find the transmitted angle θ_{t} in terms of θ_{i} (see P5.6):

$$\tan\theta_{\mathrm{t}} = \frac{n_{\mathrm{e}} \sin\theta_{\mathrm{i}}}{n_{\mathrm{o}}\sqrt{n_{\mathrm{e}}^2 - \sin^2\theta_{\mathrm{i}}}} \qquad \text{(extraordinary polarized, optic axis} \perp \text{surface)} \qquad (5.32)$$

As strange as this formula may appear, it is Snell's law, but with an angularly dependent index.

---

[4] The correspondence between s and p and ordinary and extraordinary polarization components is specific to the orientation of the optic axis in this example. For arbitrary orientations of the optic axis with respect to the surface, the ordinary and extraordinary components will generally be mixtures of s and p polarized light.

5.5 Poynting Vector in a Uniaxial Crystal

When an object is observed through a crystal (acting as a window), the energy associated with ordinary and extraordinary polarized light follow different paths, giving rise to two different images. This phenomenon is one of the more commonly observed manifestations of *birefringence*. Since the Poynting vector dictates the direction of energy flow, it is the direction of **S** that determines the separation of the double image seen when looking through a birefringent crystal.

Snell's law dictates the connection between the directions of the incident and transmitted **k**-vectors. The Poynting vector **S** for purely ordinary polarized light points in the same direction as the **k**-vector, so the direction of energy flow for ordinary polarized light also obeys Snell's law. However, for extraordinary polarized light, the Poynting vector **S** is not parallel to **k** (recall the discussion in connection with (5.5) and (5.6)). Thus, the energy flow associated with extraordinary polarized light does not obey Snell's law. When Christiaan Huygens saw this in the 1600s, one can imagine him exclaiming "how extraordinary!" Huygens' method for describing the phenomenon is outlined appendix 5.D.

To analyze extraordinary polarized light, we would like to develop an expression analogous to Snell's law, but which applies to **S** rather than to **k**. This then describes the direction that the energy associated with *extraordinary rays* takes upon entering the crystal. First, **k** inside the crystal is found from Snell's law (5.31). In general, the electric field **E** may be obtained from (5.62) and then the magnetic field via $\mathbf{B} = (\mathbf{k}\times\mathbf{E})/\omega$, to evaluate $\mathbf{S} = \mathbf{E}\times\mathbf{B}/\mu_0$. In general, this process is best done numerically, since Snell's law (5.31) for extraordinary polarized light usually does not have simple analytic solutions.

Example 5.3

Find a relationship between direction of the Poynting Vector in a uniaxial crystal and the angle of incidence in the special case where the optic axis is perpendicular to the surface.

Solution: To find the direction of energy flow, we must calculate $\mathbf{S} = \mathbf{E}\times\mathbf{B}/\mu_0$. We will need to know **E** associated with $n_e(\theta)$. We can obtain **E** from the procedures outlined in appendix 5.C. Equivalently, we can obtain it from the constitutive relation (5.3) with the definitions (5.18):

$$\begin{aligned}\epsilon_0\mathbf{E}+\mathbf{P} &= \epsilon_0\left[\left(1+\chi_x\right)E_x\hat{\mathbf{x}}+\left(1+\chi_y\right)E_y\hat{\mathbf{y}}+\left(1+\chi_z\right)E_z\hat{\mathbf{z}}\right]\\ &= \epsilon_0\left(n_o^2E_x\hat{\mathbf{x}}+n_o^2E_y\hat{\mathbf{y}}+n_e^2E_z\hat{\mathbf{z}}\right)\end{aligned} \tag{5.33}$$

Let the **k**-vector lie in the y-z plane. We may write it as $\mathbf{k} = k\left(\hat{\mathbf{y}}\sin\theta_t+\hat{\mathbf{z}}\cos\theta_t\right)$. Then the ordinary component of the field points in the x-direction, while the extraordinary component lies in the y-z plane.

Equation (5.33) requires

$$\begin{aligned}\mathbf{k}\cdot(\epsilon_0\mathbf{E}+\mathbf{P}) &= k\left(\hat{\mathbf{y}}\sin\theta_t+\hat{\mathbf{z}}\cos\theta_t\right)\cdot\epsilon_0\left(n_o^2E_x\hat{\mathbf{x}}+n_o^2E_y\hat{\mathbf{y}}+n_e^2E_z\hat{\mathbf{z}}\right)\\ &= \epsilon_0 k\left(n_o^2E_y\sin\theta_t+n_e^2E_z\cos\theta_t\right)\\ &= 0\end{aligned} \tag{5.34}$$

Therefore, the y and z components of the extraordinary field are related through

$$E_z = -\frac{n_o^2 E_y}{n_e^2} \tan\theta_t \tag{5.35}$$

We may write the extraordinary polarized electric field as

$$\mathbf{E} = E_y \left(\hat{\mathbf{y}} - \hat{\mathbf{z}} \frac{n_o^2}{n_e^2} \tan\theta_t \right) e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} \qquad \text{(extraordinary polarized)} \tag{5.36}$$

The associated magnetic field (see (2.56)) is

$$\begin{aligned} \mathbf{B} = \frac{\mathbf{k}\times\mathbf{E}}{\omega} &= \frac{k\left(\hat{\mathbf{y}}\sin\theta_t + \hat{\mathbf{z}}\cos\theta_t\right) \times E_y \left(\hat{\mathbf{y}} - \hat{\mathbf{z}}\frac{n_o^2}{n_e^2}\tan\theta_t\right)}{\omega} e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \\ &= -\hat{\mathbf{x}} \frac{kE_y}{\omega} \left(\frac{n_o^2}{n_e^2} \sin\theta_t \tan\theta_t + \cos\theta_t \right) e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \\ &\qquad \text{(extraordinary polarized)} \end{aligned} \tag{5.37}$$

The time-averaged Poynting vector then becomes

$$\begin{aligned} \langle \mathbf{S} \rangle_t &= \left\langle \mathrm{Re}\{\mathbf{E}\} \times \mathrm{Re}\left\{ \frac{\mathbf{B}}{\mu_0} \right\} \right\rangle_t \\ &= -\frac{k|E_y|^2}{\mu_0 \omega} \left(\hat{\mathbf{y}} - \hat{\mathbf{z}} \frac{n_o^2}{n_e^2} \tan\theta_t \right) \times \left(\frac{n_o^2}{n_e^2} \sin\theta_t \tan\theta_t + \cos\theta_t \right) \hat{\mathbf{x}} \left\langle \cos^2(\mathbf{k}\cdot\mathbf{r} - \omega t + \phi_{E_y}) \right\rangle_t \\ &= \frac{k|E_y|^2}{2\mu_0 \omega} \left(\frac{n_o^2}{n_e^2} \sin\theta_t \tan\theta_t + \cos\theta_t \right) \left(\hat{\mathbf{z}} + \hat{\mathbf{y}} \frac{n_o^2}{n_e^2} \tan\theta_t \right) \\ &\qquad \text{(extraordinary polarized)} \end{aligned} \tag{5.38}$$

Let us label the direction of the Poynting vector with the angle $\theta_{\mathbf{S}}$. By definition, the tangent of this angle is the ratio of the two vector components of **S**:

$$\tan\theta_{\mathbf{S}} \equiv \frac{S_y}{S_z} = \frac{n_o^2}{n_e^2} \tan\theta_t \qquad \text{(extraordinary polarized)} \tag{5.39}$$

While the **k**-vector is characterized by the angle θ_t, the Poynting vector is characterized by the angle $\theta_{\mathbf{S}}$. Combining (5.32) and (5.39), we can connect $\theta_{\mathbf{S}}$ to the incident angle θ_i:

$$\tan\theta_{\mathbf{S}} = \frac{n_o \sin\theta_i}{n_e \sqrt{n_e^2 - \sin^2\theta_i}} \qquad \text{(extraordinary polarized)} \tag{5.40}$$

As we noted in the last example, we have the case where ordinary polarized light is s-polarized light, and extraordinary polarized light is p-polarized light due to our specific choice of orientation for the optic axis in this section. In general, the s- and p-polarized portions of the incident light can each give rise to both extraordinary and ordinary rays.

Appendix 5.A Symmetry of Susceptibility Tensor

Here we show that the assumption of a nonabsorbing (and not optically active) medium implies that the susceptibility tensor is symmetric. We assume that **P** is due to a single species of electron, so that we have $\mathbf{P} = N\mathbf{p}$. Here N is the number of microscopic dipoles per volume and $\mathbf{p} = q_\mathrm{e}\mathbf{r}_\mathrm{e}$, where q_e is the charge on the electron and $\mathbf{r}_\mathrm{e}$ is the microscopic displacement of the electron. The force on this electron due to the electric field is given by $\mathbf{F} = \mathbf{E}q_\mathrm{e}$. With these definitions, we can use (5.1) to write a connection between the force due to a static **E** and the electron displacement:

$$Nq_\mathrm{e}\begin{bmatrix} x_\mathrm{e} \\ y_\mathrm{e} \\ z_\mathrm{e} \end{bmatrix} = \frac{\epsilon_0}{q_\mathrm{e}}\begin{bmatrix} \chi_{xx} & \chi_{xy} & \chi_{xz} \\ \chi_{yx} & \chi_{yy} & \chi_{yz} \\ \chi_{zx} & \chi_{zy} & \chi_{zz} \end{bmatrix}\begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} \tag{5.41}$$

The column vector on the left represents the components of the displacement $\mathbf{r}_\mathrm{e}$. We next invert (5.41) to find the force of the electric field on an electron as a function of its displacement[5]

$$\begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} = \begin{bmatrix} k_{xx} & k_{xy} & k_{xz} \\ k_{yx} & k_{yy} & k_{yz} \\ k_{zx} & k_{zy} & k_{zz} \end{bmatrix}\begin{bmatrix} x_\mathrm{e} \\ y_\mathrm{e} \\ z_\mathrm{e} \end{bmatrix} \tag{5.42}$$

where

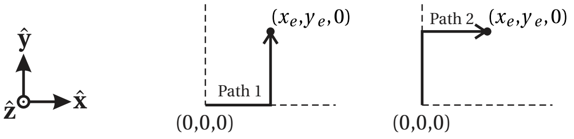
$$\begin{bmatrix} k_{xx} & k_{xy} & k_{xz} \\ k_{yx} & k_{yy} & k_{yz} \\ k_{zx} & k_{zy} & k_{zz} \end{bmatrix} \equiv \frac{Nq_\mathrm{e}^2}{\epsilon_0}\begin{bmatrix} \chi_{xx} & \chi_{xy} & \chi_{xz} \\ \chi_{yx} & \chi_{yy} & \chi_{yz} \\ \chi_{zx} & \chi_{zy} & \chi_{zz} \end{bmatrix}^{-1} \tag{5.43}$$

Here the various k_{ij} represent spring constants as opposed to components of wave vectors.

The total work done on an electron in moving it to its displaced position is given by

$$W = \int_{\text{path}} \mathbf{F}(\mathbf{r}')\cdot d\mathbf{r}' \tag{5.44}$$

While there are many possible paths for getting the electron to any specific displacement (each path specified by a different history of the electric field), the work done along any of these paths must be the same if the system is conservative (i.e. no absorption). For example, if the final displacement of $\mathbf{r} = x_\mathrm{e}\hat{\mathbf{x}} + y_\mathrm{e}\hat{\mathbf{y}}$ we could have the following two paths:



---

[5] This inversion assumes the field changes slowly so the forces on the electron are always essentially balanced. This is not true for optical fields, but the proof gives the right flavor for why conservation of energy results in the symmetry. A more formal proof that doesn't make this assumption can be found in *Principles of Optics, 7th Ed.*, Born and Wolf, pp. 790-791.

We can use (5.42) in (5.44) to calculate the total work done on the electron along path 1:

$$\begin{aligned} W &= \int_0^{x_e} F_x(x', y' = 0, z' = 0)\,dx' + \int_0^{y_e} F_y(x' = x_e, y', z' = 0)\,dy' \\ &= \int_0^{x_e} k_{xx}x'\,dx' + \int_0^{y_e} (k_{yx}x_e + k_{yy}y')\,dy' \\ &= \frac{k_{xx}}{2}x_e^2 + k_{yx}x_e y_e + \frac{k_{yy}}{2}y_e^2 \end{aligned}$$

If we take path 2, the total work is

$$\begin{aligned} W &= \int_0^{y_e} F_y(x' = 0, y', z' = 0)\,dy' + \int_0^{x_e} F_x(x', y' = y_e, z' = 0)\,dx' \\ &= \int_0^{y_e} k_{yy}y'\,dy' + \int_0^{x_e} (k_{xx}x' + k_{xy}y_e)\,dx' \\ &= \frac{k_{yy}}{2}y_e^2 + k_{xy}x_e y_e + \frac{k_{xx}}{2}x_e^2 \end{aligned}$$

Since the work must be the same for these two paths, we clearly have $k_{xy} = k_{yx}$. Similar arguments for other pairs of dimensions ensure that the matrix of k coefficients is symmetric. From linear algebra, we learn that if the inverse of a matrix is symmetric then the matrix itself is also symmetric. When we combine this result with the definition (5.43), we see that the assumption of no absorption requires the susceptibility matrix to be symmetric.

Appendix 5.B Rotation of Coordinates

In this appendix, we go through the labor of showing that (5.1) can always be written as (5.3) via rotations of the coordinate system, given that the susceptibility tensor is symmetric (i.e. $\chi_{ij} = \chi_{ji}$). We have

$$\mathbf{P} = \epsilon_0 \chi \mathbf{E} \tag{5.45}$$

where

$$\mathbf{E} \equiv \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} \qquad \mathbf{P} \equiv \begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} \qquad \chi \equiv \begin{bmatrix} \chi_{xx} & \chi_{xy} & \chi_{xz} \\ \chi_{xy} & \chi_{yy} & \chi_{yz} \\ \chi_{xz} & \chi_{yz} & \chi_{zz} \end{bmatrix} \tag{5.46}$$

Our task is to find a new coordinate system x', y', and z' for which the susceptibility tensor is diagonal. That is, we want to choose x', y', and z' such that

$$\mathbf{P}' = \epsilon_0 \chi' \mathbf{E}', \tag{5.47}$$

where

$$\mathbf{E}' \equiv \begin{bmatrix} E'_{x'} \\ E'_{y'} \\ E'_{z'} \end{bmatrix} \qquad \mathbf{P}' \equiv \begin{bmatrix} P'_{x'} \\ P'_{y'} \\ P'_{z'} \end{bmatrix} \qquad \chi' \equiv \begin{bmatrix} \chi'_{x'x'} & 0 & 0 \\ 0 & \chi'_{y'y'} & 0 \\ 0 & 0 & \chi'_{z'z'} \end{bmatrix} \tag{5.48}$$

To arrive at the new coordinate system, we are free to make pure rotation transformations. In a manner similar to (6.29), a rotation through an angle γ about the z-axis, followed by a rotation through an angle β about the resulting y-axis, and finally a rotation through an angle α about the new x-axis, can be written as

$$\begin{aligned}\mathbf{R} &\equiv \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \begin{bmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \cos\beta\cos\gamma & \cos\beta\sin\gamma & \sin\beta \\ -\cos\alpha\sin\gamma - \sin\alpha\sin\beta\cos\gamma & \cos\alpha\cos\gamma - \sin\alpha\sin\beta\sin\gamma & \sin\alpha\cos\beta \\ \sin\alpha\sin\gamma - \cos\alpha\sin\beta\cos\gamma & -\sin\alpha\cos\gamma - \cos\alpha\sin\beta\sin\gamma & \cos\alpha\cos\beta \end{bmatrix}\end{aligned} \tag{5.49}$$

The matrix $\mathbf{R}$ produces an arbitrary rotation of coordinates in three dimensions. Specifically, we can write:

$$\begin{aligned}\mathbf{E}' &= \mathbf{R}\mathbf{E} \\ \mathbf{P}' &= \mathbf{R}\mathbf{P}\end{aligned} \tag{5.50}$$

These transformations can be inverted to give

$$\begin{aligned}\mathbf{E} &= \mathbf{R}^{-1}\mathbf{E}' \\ \mathbf{P} &= \mathbf{R}^{-1}\mathbf{P}'\end{aligned} \tag{5.51}$$

where

$$\begin{aligned}\mathbf{R}^{-1} &= \begin{bmatrix} \cos\beta\cos\gamma & -\cos\alpha\sin\gamma - \sin\alpha\sin\beta\cos\gamma & \sin\alpha\sin\gamma - \cos\alpha\sin\beta\cos\gamma \\ \cos\beta\sin\gamma & \cos\alpha\cos\gamma - \sin\alpha\sin\beta\sin\gamma & -\sin\alpha\cos\gamma - \cos\alpha\sin\beta\sin\gamma \\ \sin\beta & \sin\alpha\cos\beta & \cos\alpha\cos\beta \end{bmatrix} \\ &= \begin{bmatrix} R_{11} & R_{21} & R_{31} \\ R_{12} & R_{22} & R_{32} \\ R_{13} & R_{23} & R_{33} \end{bmatrix} = \mathbf{R}^T\end{aligned} \tag{5.52}$$

Note that the inverse of the rotation matrix is the same as its transpose, an important feature that we exploit in what follows.

Upon inserting (5.51) into (5.45) we have

$$\mathbf{R}^{-1}\mathbf{P}' = \epsilon_0 \chi \mathbf{R}^{-1}\mathbf{E}' \tag{5.53}$$

or

$$\mathbf{P}' = \epsilon_0 \mathbf{R} \chi \mathbf{R}^{-1}\mathbf{E}' \tag{5.54}$$

From this equation we see that the new susceptibility tensor we seek for (5.47) is

$$\begin{aligned}\chi' &\equiv \mathbf{R}\chi\mathbf{R}^{-1}\\ &= \begin{bmatrix} R_{11} & R_{12} & R_{13}\\ R_{21} & R_{22} & R_{23}\\ R_{31} & R_{32} & R_{33}\end{bmatrix}\begin{bmatrix} \chi_{xx} & \chi_{xy} & \chi_{xz}\\ \chi_{xy} & \chi_{yy} & \chi_{yz}\\ \chi_{xz} & \chi_{yz} & \chi_{zz}\end{bmatrix}\begin{bmatrix} R_{11} & R_{21} & R_{31}\\ R_{12} & R_{22} & R_{32}\\ R_{13} & R_{23} & R_{33}\end{bmatrix}\\ &= \begin{bmatrix} \chi'_{x'x'} & \chi'_{x'y'} & \chi'_{x'z'}\\ \chi'_{x'y'} & \chi'_{y'y'} & \chi'_{y'z'}\\ \chi'_{x'z'} & \chi'_{y'z'} & \chi'_{z'z'}\end{bmatrix}\end{aligned} \tag{5.55}$$

We have expressly indicated that the off-diagonal terms of χ' are symmetric (i.e. $\chi'_{ij} = \chi'_{ji}$). This can be verified by performing the multiplication in (5.55). It is a consequence of χ being symmetric and $\mathbf{R}^{-1}$ being equal to $\mathbf{R}^T$

The three off-diagonal elements of χ' (appearing both above and below the diagonal) are found by performing the matrix multiplication in the second line of (5.55). The specific expressions for these three elements are not particularly enlightening. The important point is that we can make all three of them equal to zero since we have three degrees of freedom in the angles α, β, and γ. Although, we do not expressly solve for the angles, we have demonstrated that it is always possible to set

$$\begin{aligned}\chi'_{x'y'} &= 0\\ \chi'_{x'z'} &= 0\\ \chi'_{y'z'} &= 0\end{aligned} \tag{5.56}$$

This justifies (5.3).

Appendix 5.C Electric Field in a Crystal

To determine the direction of the electric field associated with each value of n, we return to (5.10), (5.11), and (5.12) in the analysis in section 5.2. These equations can be written in matrix format as[6]

$$\begin{bmatrix} \frac{\omega^2}{c^2}\left(1+\chi_x\right) - k_y^2 - k_z^2 & k_x k_y & k_x k_z\\ k_x k_y & \frac{\omega^2}{c^2}\left(1+\chi_y\right) - k_x^2 - k_z^2 & k_y k_z\\ k_x k_z & k_y k_z & \frac{\omega^2}{c^2}\left(1+\chi_z\right) - k_x^2 - k_y^2\end{bmatrix}\begin{bmatrix} E_x\\ E_y\\ E_z\end{bmatrix} = 0 \tag{5.57}$$

where we have used $k_x^2 + k_y^2 + k_z^2 = k^2$. We can divide every element by k^2 and employ the definitions (5.15), (5.17), and (5.18) to make this matrix equation look slightly nicer:

---

[6] A. Yariv and P. Yeh, *Optical Waves in Crystals*, Sect. 4.2 (New York: Wiley, 1984).

$$\begin{bmatrix} \frac{n_x^2}{n^2} - u_y^2 - u_z^2 & u_x u_y & u_x u_z \\ u_x u_y & \frac{n_y^2}{n^2} - u_x^2 - u_z^2 & u_y u_z \\ u_x u_z & u_y u_z & \frac{n_z^2}{n^2} - u_x^2 - u_y^2 \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} = 0 \tag{5.58}$$

For (5.58) to have a nontrivial solution (i.e. nonzero fields), the determinant of the matrix must be zero. Imposing this requirement is an equivalent way to derive Fresnel's equation (5.19) for n.

Given a direction for $\hat{\mathbf{u}}$ and a value for n (from Fresnel's equation), we can use (5.58) to determine the direction of the electric field associated with that index. It is left as an exercise to show that in nondegenerate cases[7] (i.e. $n \neq n_x, n_y, n_z$), the appropriate field direction for a value of n is given by

$$\begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} \propto \begin{bmatrix} \dfrac{u_x}{n^2 - n_x^2} \\ \dfrac{u_y}{n^2 - n_y^2} \\ \dfrac{u_z}{n^2 - n_z^2} \end{bmatrix} \qquad (n \neq n_x, n_y, n_z) \tag{5.59}$$

This is a proportionality rather than an equation because Maxwell's equations only specify the direction of **E**—we are free to choose the amplitude. Because Fresnel's equation gives two values for n, (5.59) specifies two distinct polarization components associated with each propagation direction $\hat{\mathbf{u}}$. These polarization components form a natural basis for describing light propagation in a crystal. When light is composed of a mixture of these two polarizations, the two polarization components experience different indices of refraction.

If any of the components of $\hat{\mathbf{u}}$ (i.e. u_x, u_y, or u_z) is precisely zero, the corresponding entry in (5.59) yields a zero-over-zero situation. This happens when at least one of the dimensions in (5.58) becomes decoupled from the others. In these cases, one can re-solve (5.58) for the polarization directions as in the following example.

Example 5.4

Determine the directions of the two polarization components associated with light propagating in the $\hat{\mathbf{u}} = \hat{\mathbf{z}}$ direction. (Compare with Example 5.1.)

Solution: In this case we have $u_x = u_y = 0$, so as noted above, we have to go back to (5.58) and re-solve. The set of equations becomes

$$\begin{bmatrix} \frac{n_x^2}{n^2} - 1 & 0 & 0 \\ 0 & \frac{n_y^2}{n^2} - 1 & 0 \\ 0 & 0 & \frac{n_z^2}{n^2} \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} = 0 \tag{5.60}$$

---

[7] In a biaxial crystal, the requirement $n \neq n_x, n_y, n_z$ is ensured if $u_x, u_y, u_z \neq 0$.

Notice that all three dimensions are decoupled in this system (i.e. there are no off-diagonal terms). In Example 5.1 we found that the two values of n associated with $\hat{\mathbf{u}} = \hat{\mathbf{z}}$ are n_x and n_y. If we use $n = n_x$ in our set of equations, we have

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{n_y^2}{n_x^2} - 1 & 0 \\ 0 & 0 & \frac{n_z^2}{n_x^2} \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} = 0$$

Assuming n_x and n_y are unique so that $n_y / n_x \neq 1$, these equations require $E_y = E_z = 0$ but allow E_x to be nonzero. This proves our earlier assertion that the index n_x is associated with light polarized in the x-dimension in the special case of $\hat{\mathbf{u}} = \hat{\mathbf{z}}$. Similarly, when n_y is inserted into (5.60), we find that it is associated with light polarized in the y-dimension.

We can use (5.59) to study the behavior of polarization direction as the direction of propagation varies. Figure 5.7 shows plots of the polarization direction (i.e. normalized E_x, E_y, and E_z) in potassium niobate as the propagation direction is varied. The plot is created by inserting the spherical representation of $\hat{\mathbf{u}}$ (5.24) into Fresnel's equation (5.20) for a chosen sign of the ±, and then inserting the resulting n into (5.59) to find the associated electric field. As we saw in Example 5.4, at $\theta = 0$ the light associated with the slow index is polarized along the y-axis and the light associated with the fast index is polarized along the x-axis.

In Fig. 5.7(c) we have plotted the angle between the two polarization components. At $\theta = 0$, the two polarization components are 90° apart, as you might expect. However, notice that in other propagation directions the two linear polarization components are not precisely perpendicular. Even so, the two polarization components of $\mathbf{E}$ are orthogonal in a mathematical sense,[8] so that they still comprise a useful basis for decomposing the light field.

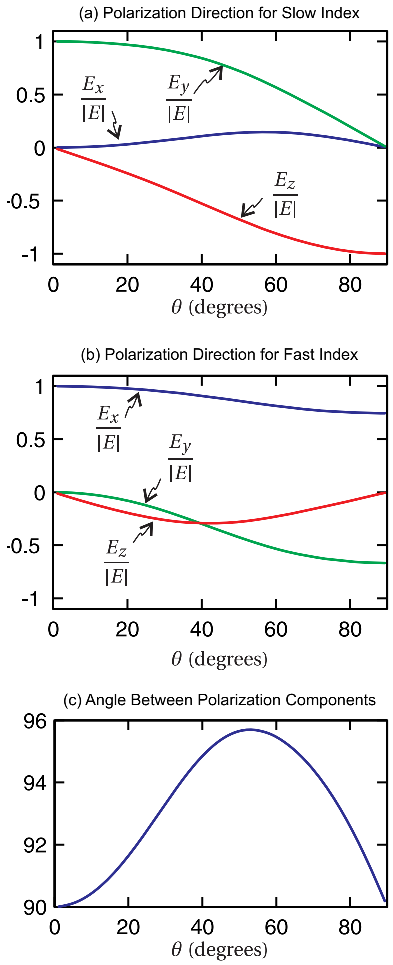


Figure 5.7 Polarization direction associated with the two values of n in potassium niobate ($KNbO_3$) at $\lambda = 500$ nm ($n_x = 2.22$, $n_y = 2.34$, and $n_z = 2.41$) and $\phi = \pi/4$. Frame (c) shows the angle between the two polarization components.

Determining the Fields in a Uniaxial Crystal.

To find the directions of the electric field for light that experiences the ordinary index of refraction in a uniaxial crystal, we insert $n = n_o$ into the requirement (5.58), and solve for the allowed fields (see P5.11) to find

$$\mathbf{E}_o(\hat{\mathbf{u}}) \propto \begin{bmatrix} -\sin\phi \\ \cos\phi \\ 0 \end{bmatrix} \tag{5.61}$$

This field component is associated with the *ordinary* wave. Just as in an isotropic medium such as glass, the index of refraction for light with this polarization does not vary with θ. The polarization component associated with $n_e(\theta)$ is found by using (5.59):

---

[8] The two components of the electric displacement vector $\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}$ remain perpendicular.

$$\mathbf{E}_\mathrm{e}(\hat{\mathbf{u}}) \propto \begin{bmatrix} \dfrac{\sin\theta\cos\phi}{n_\mathrm{e}^2(\theta) - n_\mathrm{o}^2} \\[2ex] \dfrac{\sin\theta\sin\phi}{n_\mathrm{e}^2(\theta) - n_\mathrm{o}^2} \\[2ex] \dfrac{\cos\theta}{n_\mathrm{e}^2(\theta) - n_\mathrm{e}^2} \end{bmatrix} \tag{5.62}$$

Notice that this polarization component is partially directed along the optic axis (i.e. it has a z-component), and it is *not* perpendicular to $\mathbf{k}$ since $\hat{\mathbf{u}} \cdot \mathbf{E}_\mathrm{e}(\hat{\mathbf{u}}) \neq 0$ (see P5.12). It is, however, perpendicular to the ordinary polarization component, since $\mathbf{E}_\mathrm{e} \cdot \mathbf{E}_\mathrm{o} = 0$.

Notice that when $\theta = 0$, (5.29) reduces to $n = n_\mathrm{o}$ so that both indices are the same. On the other hand, if $\theta = \pi/2$ then (5.29) reduces to $n = n_\mathrm{e}$. These limits must be approached carefully in (5.62).

Appendix 5.D Huygens' Elliptical Construct for a Uniaxial Crystal

In 1690 Christian Huygens developed a way to predict the direction of extraordinary rays in a crystal by examining an elliptical wavelet. The point on the elliptical wavelet that propagates along the optic axis is assumed to experience the index n_e. The point on the elliptical wavelet that propagates perpendicular to the optic axis is assumed to experience the index n_o. It turns out that Huygens' approach agreed with the direction energy propagation (5.40) (as opposed to the direction of the **k**-vector). This was quite satisfactory in Huygens' day (except that he was largely ignored for a century, owing to Newton's corpuscular theory) since the direction of energy propagation is what an observer sees.

Consider a plane wave entering a uniaxial crystal with the optic axis perpendicular to the surface. In Huygens' point of view, each point on a wavefront acts as a wavelet source which combines with neighboring wavelets to preserve the overall plane wave pattern. Inside the crystal, the wavelets propagate in the shape of an ellipse. The equation for an elliptical wave front after propagating during a time t is

$$\frac{y^2}{(ct/n_\mathrm{e})^2} + \frac{z^2}{(ct/n_\mathrm{o})^2} = 1 \tag{5.63}$$

After rearranging, the equation of the ellipse can be written as

$$z = \frac{ct}{n_\mathrm{o}}\sqrt{1 - \frac{y^2}{(ctn_\mathrm{e})^2}} \tag{5.64}$$

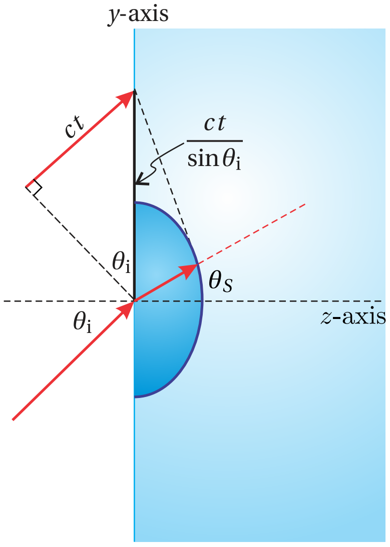


Figure 5.8 Elliptical wavelet.

In order to have the wavelet joint neatly with other wavelets to build a plane wave, the wavefront of the ellipse must be parallel to a new wavefront entering the surface at a distance $ct/\sin\theta_\mathrm{i}$ above the original point. This distance is represented by the hypotenuse of the right triangle seen in Fig. 5.8. Let the point where the wavefront touches the ellipse be denoted by $\left(y,z\right) = (z\tan\theta_{\mathbf{S}}, z)$. The slope (rise over run) of the line that connects these two points is then

$$\frac{dz}{dy} = -\frac{z}{ct/\sin\theta_\mathrm{i} - z\tan\theta_{\mathbf{S}}} \tag{5.65}$$

At the point where the wavefront touches the ellipse (i.e. $\left(y,z\right) = (z\tan\theta_{\mathbf{S}}, z)$), the slope of the curve for the ellipse is

$$\frac{dz}{dy} = \frac{-yn_\mathrm{e}^2}{n_\mathrm{o}ct\sqrt{1-\frac{y^2}{(ct/n_\mathrm{e})^2}}} = -\frac{n_\mathrm{e}^2 y}{n_\mathrm{o}^2 z} = -\frac{n_\mathrm{e}^2}{n_\mathrm{o}^2}\tan\theta_{\mathbf{S}} \tag{5.66}$$

We would like these two slopes to be the same. We therefore set them equal to each other:

$$-\frac{n_\mathrm{e}^2}{n_\mathrm{o}^2}\tan\theta_{\mathbf{S}} = -\frac{z}{ct/\sin\theta_\mathrm{i} - z\tan\theta_{\mathbf{S}}} \Rightarrow \frac{ct}{z}\frac{n_\mathrm{e}^2}{n_\mathrm{o}^2}\frac{\tan\theta_{\mathbf{S}}}{\sin\theta_\mathrm{i}} = \frac{n_\mathrm{e}^2}{n_\mathrm{o}^2}\tan^2\theta_{\mathbf{S}} + 1 \tag{5.67}$$

If we evaluate (5.63) for the point $\left(y,z\right) = (z\tan\theta_{\mathbf{S}}, z)$, we obtain

$$\frac{ct}{z} = n_\mathrm{o}\sqrt{\frac{n_\mathrm{e}^2}{n_\mathrm{o}^2}\tan^2\theta_{\mathbf{S}} + 1} \tag{5.68}$$

Upon substitution of this into (5.67) we arrive at

$$\frac{n_\mathrm{e}^2}{n_\mathrm{o}}\frac{\tan\theta_{\mathbf{S}}}{\sin\theta_\mathrm{i}} = \sqrt{\frac{n_\mathrm{e}^2}{n_\mathrm{o}^2}\tan^2\theta_{\mathbf{S}} + 1} \Rightarrow \frac{n_\mathrm{e}^4}{n_\mathrm{o}^2}\frac{\tan^2\theta_{\mathbf{S}}}{\sin^2\theta_\mathrm{i}} = \frac{n_\mathrm{e}^2}{n_\mathrm{o}^2}\tan^2\theta_{\mathbf{S}} + 1 \tag{5.69}$$

$$\Rightarrow \left[\frac{n_\mathrm{e}^2}{\sin^2\theta_\mathrm{i}} - 1\right]\tan^2\theta_{\mathbf{S}} = \frac{n_\mathrm{o}^2}{n_\mathrm{e}^2} \Rightarrow \tan\theta_{\mathbf{S}} = \frac{n_\mathrm{o}\sin\theta_\mathrm{i}}{n_\mathrm{e}\sqrt{n_\mathrm{e}^2 - \sin^2\theta_\mathrm{i}}} \tag{5.70}$$

This agrees with (5.40) as anticipated. Again, Huygens' approach obtained the correct direction of the Poynting vector associated with the extraordinary wave.

Exercises

Exercises for 5.2 Plane Wave Propagation in Crystals

P5.1 (a) Solve Fresnel's equation (5.19) to find the two values of n^2 associated with a given $\hat{\mathbf{u}}$. In other words, fill in the steps leading to (5.20)–(5.23).

(b) Point out that both solutions for n^2 are real and positive, when n_x, n_y, and n_z are real and $B^2 - 4AC \geq 0$ in (5.20). Show that $B^2 - 4AC \geq 0$ in the following special cases: Case I: $u_x = \pm 1$, $u_y = 0$, $u_z = 0$; Case II: $u_x = \pm 1/\sqrt{3}$, $u_y = \pm 1/\sqrt{3}$, $u_z = \pm 1/\sqrt{3}$.

HINT: First manipulate (5.19) into the form

$$
\begin{aligned}
&\left[\left(u_x^2 + u_y^2 + u_z^2\right) - 1\right] n^6 \\
&+ \left[\left(n_x^2 + n_y^2 + n_z^2\right) - u_x^2\left(n_y^2 + n_z^2\right) - u_y^2\left(n_x^2 + n_z^2\right) - u_z^2\left(n_x^2 + n_y^2\right)\right] n^4 \\
&- \left[\left(n_x^2 n_y^2 + n_x^2 n_z^2 + n_y^2 n_z^2\right) - u_x^2 n_y^2 n_z^2 - u_y^2 n_x^2 n_z^2 - u_z^2 n_x^2 n_y^2\right] n^2 + n_x^2 n_y^2 n_z^2 = 0
\end{aligned}
$$

Substitute $1 = u_x^2 + u_y^2 + u_z^2$ in several places.

P5.2 Show that Fresnel's equation (5.19) may equivalently be written as.

$$
\frac{u_x^2}{\left(\frac{1}{n^2} - \frac{1}{n_x^2}\right)} + \frac{u_y^2}{\left(\frac{1}{n^2} - \frac{1}{n_y^2}\right)} + \frac{u_z^2}{\left(\frac{1}{n^2} - \frac{1}{n_z^2}\right)} = 0
$$

HINT: Use $1 = u_x^2 + u_y^2 + u_z^2$.

P5.3 Suppose you have a crystal with $n_x = 1.5$, $n_y = 1.6$, and $n_z = 1.7$. Use Fresnel's equation to determine what the two indices of refraction are for a k-vector in the crystal along the $\hat{\mathbf{u}} = (\hat{\mathbf{x}} + 2\hat{\mathbf{y}} + 3\hat{\mathbf{z}})/\sqrt{14}$ direction.

Exercises for 5.3 Biaxial and Uniaxial Crystals

P5.4 (a) Show that for a biaxial crystal, the directions of the optic axes are given by (5.25) in the x-z plane.

(b) Show that (5.25) only makes sense if the axes are chosen such that n_y is in between n_x and n_z. Does the formula work if $n_x \geq n_y \geq n_z$? How do the values of θ relate to the case when $n_z \geq n_y \geq n_x$?

HINT: Use spherical coordinates as in (5.24). The two indices are the same when $B^2 - 4AC = 0$. Under the assumption that n_y lies between n_x and n_z, $B^2 - 4AC = 0$ can only be satisfied when $\phi = 0$.

P5.5 Use definitions (5.26) and (5.27) along with the spherical representation of $\hat{\mathbf{u}}$ (5.24) in Fresnel's equation (5.20) to calculate the two values for the index in a uniaxial crystal (i.e. (5.28) and (5.29)).

HINT: First show that

$$\begin{aligned} A &= n_\text{o}^2 \sin^2\theta + n_\text{e}^2\cos^2\theta \\ B &= n_\text{o}^2 n_\text{e}^2 + n_\text{o}^4 \sin^2\theta + n_\text{e}^2 n_\text{o}^2\cos^2\theta \\ C &= n_\text{o}^4 n_\text{e}^2 \end{aligned}$$

and then use these expressions to evaluate Fresnel's equation.

Exercises for 5.4 Refraction at a Uniaxial Crystal Surface

P5.6 Derive (5.32).

P5.7 Suppose you have a quartz plate (a uniaxial crystal) with its optic axis oriented perpendicular to the surfaces. The indices of refraction for quartz are $n_\text{o} = 1.54424$ and $n_\text{e} = 1.55335$. A plane wave with wavelength $\lambda_\text{vac} = 633$ nm passes through the plate. After emerging from the crystal, there is a phase difference $\Delta\phi$ between the two polarization components of the plane wave, and this phase difference depends on incident angle θ_i. Use a computer to plot $\Delta\phi$ as a function of incident angle from zero to 90° for a plate with thickness $d = 0.96$ mm .

HINT: For s-polarized light, show that the number of wavelengths that fit in the plate is $\frac{d}{(\lambda_\text{vac}/n_\text{o})\cos\theta_t^{(s)}}$. For p-polarized light, show that the number of wavelengths that fit in the plate and the extra leg δ outside of the plate (see Fig. 5.9) is $\frac{d}{(\lambda_\text{vac}/n_p)\cos\theta_t^{(p)}} + \frac{\delta}{\lambda_\text{vac}}$, where

$$\delta = d\left[\tan\theta_t^{(s)} - \tan\theta_t^{(p)}\right]\sin\theta_\text{i}$$

and n_p is given by (5.29). Find the difference between these expressions and multiply by 2π to find $\Delta\phi$.

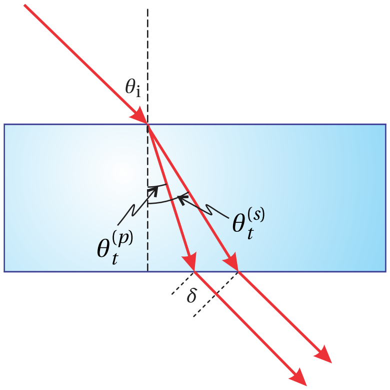


Figure 5.9 Diagram for P5.7.

L5.8 In the laboratory, send a HeNe laser ($\lambda_\text{vac} = 633$ nm) through two crossed polarizers, oriented at 45° and 135°. Place the quartz plate described in P5.7 between the polarizers on a rotation stage. Now equal amounts of s- and p-polarized light strike the crystal as it is rotated from normal incidence. (video)

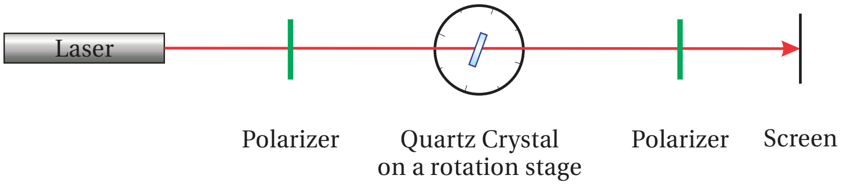


Figure 5.11 Schematic for L 5.8.

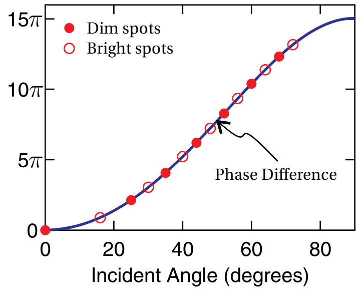


Figure 5.10 Plot for P5.7 and L 5.8.

If the phase shift between the two paths discussed in P5.7 is an odd integer times π, the polarization direction of the light transmitted through the crystal is rotated by 90°, and the maximum transmission through the second polarizer results. (In this configuration, the crystal acts as a *half-wave plate*, which we discuss in Chapter 6.) If the phase shift is an even integer times π, then the polarization is rotated by 180° and minimum transmission through the second polarizer results. Plot these measured maximum and minimum points on your computer-generated graph of the previous problem.

Exercises for 5.5 Poynting Vector in a Uniaxial Crystal

P5.9 A calcite crystal is cut and polished such that the optic axis is perpendicular to the surface.[9] If 590 nm light enters with incident angle $\theta_\mathrm{i} = 45°$, what is the difference between the transmitted angles of the Poynting vector for s- and p-polarized light? Calcite is a uniaxial crystal with $n_\mathrm{o} = 1.658$ and $n_\mathrm{e} = 1.486$ at this wavelength.

Exercises for 5.C Electric Field in a Crystal

P5.10 Check that (5.59) is a solution to (5.58).

P5.11 (a) Show that the field polarization component associated with $n = n_\mathrm{o}$ in a uniaxial crystal is given by (5.61)by substituting this value for n into (5.58) and determining what combination of field components are allowable.

(b) Show that the field is directed perpendicular to the plane containing $\hat{\mathbf{u}}$ and $\hat{\mathbf{z}}$.

P5.12 (a) Show that the electric field for extraordinary polarized light $\mathbf{E}_\mathrm{e}(\hat{\mathbf{u}})$ in a uniaxial crystal is not perpendicular to $\mathbf{k}$ (i.e. $\hat{\mathbf{u}}$).

(b) Show that the ordinary polarization component $\mathbf{E}_\mathrm{o}(\hat{\mathbf{u}})$ is perpendicular to $\mathbf{k}$.

---

[9] This is called an a-cut. Calcite cleaves naturally along its rhombohedron form, which is not the same as an a-cut.

Chapter 6

Polarization of Light

When the direction of the electric field of light oscillates in a regular, predictable fashion, we say that the light is *polarized*. *Polarization* describes the direction of the oscillating electric field, a distinct concept from dipoles per volume in a material **P** – also called polarization. In this chapter, we develop a formalism for describing polarized light and the effect of devices that modify polarization. If the electric field oscillates in a plane, we say that it is *linearly polarized*. The electric field can also spiral around while a plane wave propagates, and this is called *circular* or *elliptical polarization*. There is a convenient way for keeping track of polarization using a two-dimensional *Jones vector*.

Many devices can affect polarization such as *polarizers* and *wave plates*. Their effects on a light field can be represented by 2 × 2 *Jones matrices* that operate on the Jones vector representing the light. A Jones matrix can describe, for example, a polarizer oriented at an arbitrary angle or it can characterize the influence of a wave plate, which is a device that introduces a relative phase between two components of the electric field.

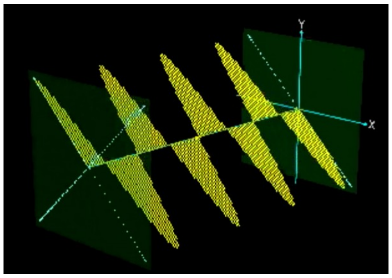


Figure 6.1 Animation showing different polarization states of light.

In this chapter, we will also see how reflection and transmission at a material interface influences field polarization. As we saw previously, s-polarized light can acquire a phase lag or phase advance relative to p-polarized light. This is especially true at metal surfaces, which have complex indices of refraction. The Fresnel coefficients studied in chapters 3 and 4 can be conveniently incorporated into a Jones matrix to keep track of their influence polarization. *Ellipsometry*, outlined in appendix 6.A, is the science of characterizing optical properties of materials through an examination of these effects.

Throughout this chapter, we consider light to have well characterized polarization. However, most common sources of light (e.g. sunlight or a light bulb) have an electric-field direction that varies rapidly and randomly. Such sources are commonly referred to as *unpolarized*. It is common to have a mixture of unpolarized and polarized light, called *partially polarized* light. The Jones vector formalism used in this chapter is inappropriate for describing the unpolarized portions of the light. In appendix 6.B we describe a more general formalism for dealing with light having an arbitrary *degree of polarization*.

6.1 Linear, Circular, and Elliptical Polarization

Consider the plane-wave solution to Maxwell's equations given by

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{6.1}$$

The wave vector **k** specifies the direction of propagation. We neglect absorption so that the refractive index is real and $k = n\omega/c = 2\pi n/\lambda_{\text{vac}}$ (see (2.19)–(2.24)). In an isotropic medium we know that **k** and $\mathbf{E}_0$ are perpendicular, but even after the direction of **k** is specified, we are still free to have $\mathbf{E}_0$ point anywhere in the two dimensions perpendicular to **k**. If we orient our coordinate system with the z-axis in the direction of **k**, we can write (6.1) as

$$\mathbf{E}(z,t) = \left(E_x\hat{\mathbf{x}} + E_y\hat{\mathbf{y}}\right) e^{i(kz-\omega t)} \tag{6.2}$$

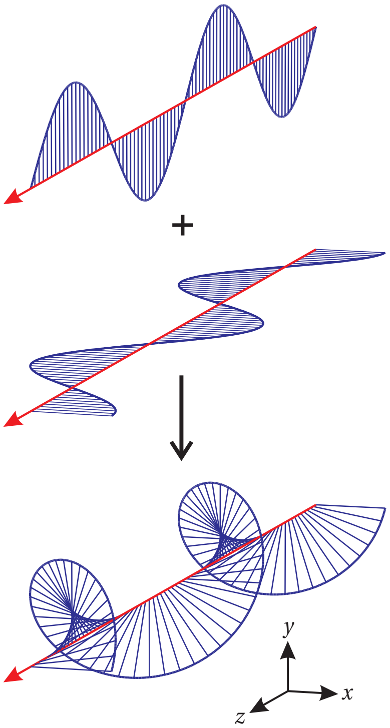


Figure 6.2 The combination of two orthogonally polarized plane waves that are out of phase results in elliptically polarized light. Here we have left circularly polarized light created as specified by (6.3).

As always, only the real part of (6.2) is physically relevant. The complex amplitudes E_x and E_y keep track of the phase of the oscillating field components. In general the complex phases of E_x and E_y can differ, so that the wave in one of the dimensions lags or leads the wave in the other dimension.

The relationship between E_x and E_y describes the polarization of the light. For example, if E_y is zero, the plane wave is said to be *linearly polarized* along the x-dimension. Linearly polarized light can have any orientation in the x–y plane, and it occurs whenever E_x and E_y have the same complex phase (or phases differing by π). For our purposes, we will take the x-dimension to be horizontal and the y-dimension to be vertical unless otherwise noted.

As an example, suppose $E_y = iE_x$, where E_x is real. The y-component of the field is then out of phase with the x-component by the factor $i = e^{i\pi/2}$. Taking the real part of the field (6.2) we get

$$\begin{aligned}\mathbf{E}(z,t) &= \mathrm{Re}\left[E_x e^{i(kz-\omega t)}\right]\hat{\mathbf{x}} + \mathrm{Re}\left[e^{i\pi/2}E_x e^{i(kz-\omega t)}\right]\hat{\mathbf{y}} \\ &= E_x\cos(kz-\omega t)\,\hat{\mathbf{x}} + E_x\cos(kz-\omega t+\pi/2)\,\hat{\mathbf{y}} \qquad \text{(left circular)} \\ &= E_x\left[\cos(kz-\omega t)\,\hat{\mathbf{x}} - \sin(kz-\omega t)\,\hat{\mathbf{y}}\right]\end{aligned} \tag{6.3}$$

In this example, the field in the y-dimension lags in time behind the field in the x-dimension by a quarter cycle. That is, the behavior seen in the x-dimension happens in the y-dimension a quarter cycle later. The field never goes to zero simultaneously in both dimensions. In fact, in this example the strength of the electric field is constant, and it rotates in a circular pattern in the x-y dimensions. For this reason, this type of field is called *circularly polarized*. Figure 6.2 graphically shows the two linear polarized pieces in (6.3) adding to make circularly polarized light.

If we view a circularly polarized light field throughout space at a frozen instant in time (as in Fig. 6.2), the electric field vector spirals as we move along the z-dimension. If the sense of the spiral (with time frozen) matches that of a common wood screw oriented along the z-axis, the polarization is called *right handed*. (It makes no difference whether the screw is flipped end for end.) If instead the field spirals in the opposite sense, then the polarization is called *left handed.* The field shown in Fig. 6.2 is an example of left-handed circularly polarized light.

An equivalent way to view the handedness convention is to imagine the light impinging on a screen as a function of time. The field of a right-handed circularly polarized wave rotates counterclockwise at the screen, when looking along the **k** direction. The field rotates clockwise for a left-handed circularly polarized wave.

Linearly polarized light can become circularly or, in general, *elliptically* polarized after reflection from a metal surface if the incident light has both s- and p-polarized components. A good experimentalist working with light needs to know this. Reflections from multilayer dielectric mirrors can also exhibit these phase shifts.

6.2 Jones Vectors for Representing Polarization

In 1941, R. Clark Jones introduced a two-dimensional matrix algebra that is useful for keeping track of light polarization and the effects of optical elements that influence polarization.[1] The algebra deals with light having a definite polarization, such as plane waves. It does not apply to unpolarized or partially polarized light (e.g. sunlight). For partially polarized light, a four-dimensional algebra known as Stokes calculus is used (see Appendix 6.B).

In preparation for introducing Jones vectors, we explicitly write the complex phases of the field components in (6.2) as

$$\mathbf{E}(z,t) = \left(|E_x|e^{i\phi_x}\hat{\mathbf{x}} + |E_y|e^{i\phi_y}\hat{\mathbf{y}}\right)e^{i(kz-\omega t)} \tag{6.4}$$

and then factor (6.4) as follows:

$$\mathbf{E}(z,t) = E_{\text{eff}}\left(A\hat{\mathbf{x}} + Be^{i\delta}\hat{\mathbf{y}}\right)e^{i(kz-\omega t)} \tag{6.5}$$

where

$$E_{\text{eff}} \equiv \sqrt{|E_x|^2 + \left|E_y\right|^2}\,e^{i\phi_x} \tag{6.6}$$

$$A \equiv \frac{|E_x|}{\sqrt{|E_x|^2 + \left|E_y\right|^2}} \tag{6.7}$$

$$B \equiv \frac{\left|E_y\right|}{\sqrt{|E_x|^2 + \left|E_y\right|^2}} \tag{6.8}$$

$$\delta \equiv \phi_y - \phi_x \tag{6.9}$$

Please notice that A and B are real nonnegative dimensionless numbers that satisfy $A^2 + B^2 = 1$. If E_y is zero, then $B = 0$ and everything is well-defined. On the other hand, if E_x happens to be zero, then its phase $e^{i\phi_x}$ is indeterminate. In this case we let $E_{\text{eff}} = |E_y|e^{i\phi_y}$, $B = 1$, and $\delta = 0$.

R. Clark Jones (1916–2004, American) was born in Toledo Ohio. He was one of six high school seniors to receive a Harvard College National Prize Fellowship. He earned both his undergraduate (summa cum laude 1938) and Ph.D. degrees from Harvard (1941). After working several years at Bell Labs, he spent most of his professional career at Polaroid Corporation in Cambridge MA, until his retirement in 1982. He is well-known for a series of papers on polarization published during the period 1941-1956. He also contributed greatly to the development of infrared detectors. He was an avid train enthusiast, and even wrote papers on railway engineering. See J. Opt. Soc. Am. **63**, 519-522 (1972). Also see SPIE oemagazine, p. 52 (Aug. 2004).

---

[1] E. Hecht, *Optics*, 3rd ed., Sect. 8.12.2 (Massachusetts: Addison-Wesley, 1998).

Linearly polarized along x

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Linearly polarized along y

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Linearly polarized at angle α (measured from the x-axis)

$$\begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix}$$

Right circularly polarized

$$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -i \end{bmatrix}$$

Left circularly polarized

$$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ i \end{bmatrix}$$

Table 6.1 Jones Vectors for several common polarization states.

The overall field strength E_{eff} is often unimportant in a discussion of polarization. It represents the strength of an *effective* linearly polarized field that would correspond to the same intensity as (6.4). Specifically, from (2.62) and (6.5) we have

$$I = \langle S \rangle_t = \frac{1}{2} n c \epsilon_0 \mathbf{E} \cdot \mathbf{E}^* = \frac{1}{2} n c \epsilon_0 \left| E_{\text{eff}} \right|^2 \tag{6.10}$$

The phase of E_{eff} represents an overall phase shift that one can trivially adjust by physically moving the light source (a laser, say) forward or backward by a fraction of a wavelength.

The portion of (6.5) that is relevant to our discussion of polarization is the vector $A\hat{\mathbf{x}} + Be^{i\delta}\hat{\mathbf{y}}$, referred to as the *Jones vector*. This vector contains the essential information regarding field polarization. Notice that the Jones vector is a kind of unit vector, in that $(A\hat{\mathbf{x}} + Be^{i\delta}\hat{\mathbf{y}}) \cdot (A\hat{\mathbf{x}} + Be^{i\delta}\hat{\mathbf{y}})^* = 1$. (The asterisk represents the complex conjugate.) When writing a Jones vector we dispense with the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ notation and organize the components into a column vector (for later use in matrix algebra) as follows:

$$\begin{bmatrix} A \\ Be^{i\delta} \end{bmatrix} \tag{6.11}$$

This vector can describe the polarization state of any plane wave field. Table 6.1 lists some Jones vectors representing various polarization states.

6.3 Elliptically Polarized Light

In general, the Jones vector (6.11) represents a polarization state between linear and circular. This 'between' state is known as *elliptically polarized* light. As the wave travels, the field vector makes a spiral motion. If we observe the field vector at a point as the field goes by, the field vector traces out an ellipse oriented perpendicular to the direction of travel (i.e. in the x–y plane). One of the axes of the ellipse occurs at the angle

$$\alpha = \frac{1}{2} \tan^{-1}\left(\frac{2AB\cos\delta}{A^2 - B^2} \right) \tag{6.12}$$

with respect to the x-axis (see P6.8). This angle sometimes corresponds to the minor axis and sometimes to the major axis of the ellipse, depending on the exact values of A, B, and δ. The other axis of the ellipse (major or minor) then occurs at $\alpha \pm \pi/2$ (see Fig. 6.3).

We can deduce whether (6.12) corresponds to the major or minor axis of the ellipse by comparing the strength of the electric field when it spirals through the direction specified by α and when it spirals through $\alpha \pm \pi/2$. The strength of the electric field at α is given by (see P6.8)

$$E_\alpha = \left| E_{\text{eff}} \right| \sqrt{A^2 \cos^2\alpha + B^2 \sin^2\alpha + AB\cos\delta \sin 2\alpha} \qquad (E_{\max} \text{ or } E_{\min}) \tag{6.13}$$

and the strength of the field when it spirals through the orthogonal direction ($\alpha \pm \pi/2$) is given by

$$E_{\alpha\pm\pi/2} = |E_{\text{eff}}|\sqrt{A^2\sin^2\alpha + B^2\cos^2\alpha - AB\cos\delta\sin 2\alpha} \qquad (E_{\text{max}} \text{ or } E_{\text{min}}) \tag{6.14}$$

After computing (6.13) and (6.14), we decide which represents E_{min} and which E_{max} according to

$$E_{\text{max}} \geq E_{\text{min}} \tag{6.15}$$

We could predict in advance which of (6.13) or (6.14) corresponds to the major axis and which corresponds to the minor axis. However, making this prediction is as complicated as simply evaluating (6.13) and (6.14) and determining which is greater.

Elliptically polarized light is often characterized by the *ellipticity*, given by the ratio of the minor axis to the major axis:

$$e \equiv \frac{E_{\text{min}}}{E_{\text{max}}} \tag{6.16}$$

The ellipticity e ranges between zero (corresponding to linearly polarized light) and one (corresponding to circularly polarized light). Finally, the *helicity* or handedness of elliptically polarized light is as follows (see P6.2):

$$0 < \delta < \pi \quad \rightarrow \quad \text{left-handed helicity} \tag{6.17}$$

$$\pi < \delta < 2\pi \quad \rightarrow \quad \text{right-handed helicity} \tag{6.18}$$

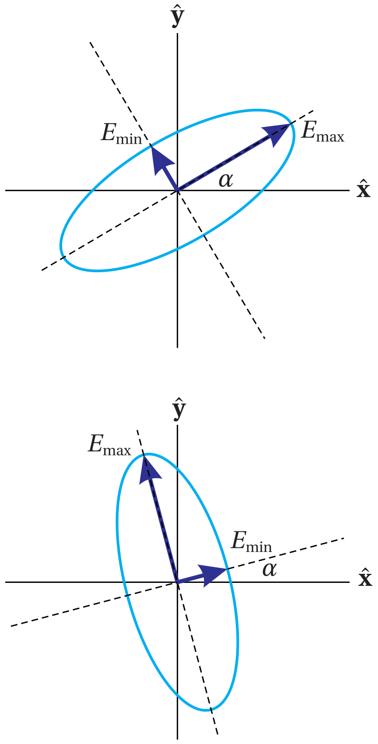


Figure 6.3 The electric field of elliptically polarized light traces an ellipse in the plane perpendicular to its propagation direction. The two plots are for different values of A, B, and δ. The angle α can describe the major axis (top) or the minor axis (bottom), depending on the values of these parameters.

6.4 Linear Polarizers and Jones Matrices

In 1928, Edwin Land invented an inexpensive polarizing device. He did it by stretching a polymer sheet and infusing it with iodine. The stretching caused the polymer chains to align along a common direction, whereupon the sheet was cemented to a substrate. The infusion of iodine caused the individual chains to become conductive, like microscopic wires.

When light impinges upon Land's Polaroid sheet, the component of electric field that is *parallel* to the polymer chains causes a current $\mathbf{J}_{\text{free}}$ to oscillate in that dimension. The resistance to the current quickly dissipates the energy (i.e. the refractive index is complex) and the light is absorbed. The thickness of the Polaroid sheet is chosen sufficiently large to ensure that virtually none of the light with electric field component oscillating along the chains makes it through the device.

The component of electric field that is orthogonal to the polymer chains encounters electrons that are essentially bound to the narrow width of individual polymer molecules. For this polarization component, the wave passes through the material much like it does through typical dielectrics such as glass (i.e. the refractive index is real). Today, there is a wide variety of technologies for making polarizers, many very different from Polaroid.

A polarizer can be represented as a 2×2 matrix that operates on Jones vectors.[2] The function of a polarizer is to pass only the component of electric field that is oriented along the polarizer transmission axis. If a polarizer is oriented with its transmission axis along the x-dimension, only the x-component of polarization transmits; the y-component is killed. If the polarizer is oriented with its transmission axis along the y-dimension, only the y-component of the field transmits, and the x-component is killed. These two scenarios can be represented with the following Jones matrices:

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad \text{(polarizer with transmission along x-axis)} \tag{6.19}$$

$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \qquad \text{(polarizer with transmission along y-axis)} \tag{6.20}$$

These matrices operate on any Jones vector representing the polarization of incident light. The result gives the Jones vector for the light exiting the polarizer.

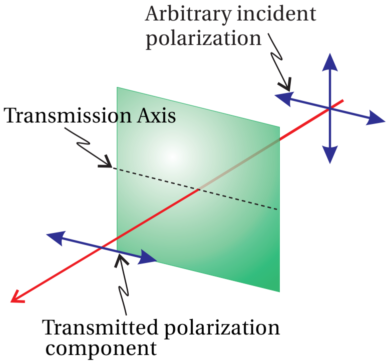

Arbitrary incident polarization
Transmission Axis
Transmitted polarization component

Figure 6.4 Light transmitting through a Polaroid sheet. The conducting polymer chains run vertically in this drawing, and light polarized along the chains is absorbed. Light polarized perpendicular to the polymer chains passes through the polarizer.

Example 6.1

Use the Jones matrix (6.19) to calculate the effect of a horizontal polarizer on light that is initially horizontally polarized, vertically polarized, and arbitrarily polarized.

Solution: First we consider a horizontally polarized plane wave traversing a polarizer with its transmission axis oriented also horizontally (x-dimension):

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \text{(horizontal polarizer on horizontally polarized field)}$$

As expected, the polarization state is unaffected by the polarizer. (We have ignored possible attenuation from surface reflections.)

Now consider vertically polarized light traversing the same horizontal polarizer. In this case, we have:

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \text{(horizontal polarizer on vertical linear polarization)}$$

As expected, the polarizer extinguishes the light.

Finally, when a horizontally oriented polarizer operates on light with an arbitrary Jones vector (6.11), we have

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} A \\ Be^{i\delta} \end{bmatrix} = \begin{bmatrix} A \\ 0 \end{bmatrix} \qquad \text{(horizontal polarizer on arbitrary polarization)}$$

Only the horizontal component of polarization is transmitted through the polarizer.

---

[2] E. Hecht, *Optics*, 3rd ed., Sect. 8.12.3 (Massachusetts: Addison-Wesley, 1998).

While you might readily agree that the matrices given in (6.19) and (6.20) can be used to get the right result for light traversing a horizontal or a vertical polarizer, you probably aren't very impressed as of yet. In the next few sections, we will derive Jones matrices for a number of optical elements that can modify polarization: polarizers at arbitrary angle, wave plates at arbitrary angle, and reflection or transmissions at an interface. Table 6.2 shows Jones matrices for each of these devices. Before deriving these specific Jones matrices, however, we take a moment to appreciate why the Jones matrix formulation is useful.

Linear polarizer

$$\begin{bmatrix} \cos^2\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & \sin^2\theta \end{bmatrix}$$

Half-wave plate

$$\begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix}$$

Quarter-wave plate

$$\begin{bmatrix} \cos^2\theta + i\sin^2\theta & (1-i)\sin\theta\cos\theta \\ (1-i)\sin\theta\cos\theta & \sin^2\theta + i\cos^2\theta \end{bmatrix}$$

Right circular polarizer

$$\frac{1}{2}\begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix}$$

Left circular polarizer

$$\frac{1}{2}\begin{bmatrix} 1 & -i \\ i & 1 \end{bmatrix}$$

Reflection from an interface

$$\begin{bmatrix} -r_p & 0 \\ 0 & r_s \end{bmatrix}$$

Transmission through an interface

$$\begin{bmatrix} t_p & 0 \\ 0 & t_s \end{bmatrix}$$

Table 6.2 Common Jones Matrices. The angle θ is measured with respect to the x-axis and specifies the transmission axis of a linear polarizer or the fast axis of a wave plate.

The real power of the formalism becomes clear as we consider situations where light encounters multiple polarization elements in sequence. In these situations, we use a product of Jones matrices to represent the effect of the compound systems. We can represent this situation by

$$\begin{bmatrix} A' \\ B' \end{bmatrix} = \mathbf{J}_{\text{system}} \begin{bmatrix} A \\ Be^{i\delta} \end{bmatrix} \tag{6.21}$$

where the unprimed Jones vector represents light going into the system and the primed Jones vector represents light emerging from the system.

The matrix $\mathbf{J}_{\text{system}}$ is a Jones matrix formed by a series of polarization devices. If there are N devices in the system, the compound matrix is calculated as

$$\mathbf{J}_{\text{system}} \equiv \mathbf{J}_N \mathbf{J}_{N-1} \cdots \mathbf{J}_2 \mathbf{J}_1 \tag{6.22}$$

where $\mathbf{J}_n$ is the matrix for the n^{th} polarizing optical element encountered in the system. Notice that the matrices operate on the Jones vector in the order that the light encounters the devices. Therefore, the matrix for the first device ($\mathbf{J}_1$) is written on the *right*, and so on until the last device encountered, which is written on the *left*, farthest from the Jones vector.

When part of the light is absorbed by passing through one or more polarizers in a system, the Jones vector of the exiting light does not necessarily remain normalized to magnitude one (see Example 6.1). The factor by which the intensity of the light decreases is given by $\left(A'\hat{\mathbf{x}} + B'\hat{\mathbf{y}}\right) \cdot \left(A'\hat{\mathbf{x}} + B'\hat{\mathbf{y}}\right)^* = \left|A'\right|^2 + \left|B'\right|^2$. The intensity exiting from the system is then

$$I' = \frac{1}{2} n c \epsilon_0 \left|E'_{\text{eff}}\right|^2 \quad \text{where} \quad \left|E'_{\text{eff}}\right|^2 = \left|E_{\text{eff}}\right|^2 \left(\left|A'\right|^2 + \left|B'\right|^2\right) \tag{6.23}$$

Here, E_{eff} is the original effective field before entering the system (see (6.10)), and E'_{eff} is the final effective field.

For the sake of further analysis, if desired, one can renormalize the final Jones vector and write it in standard form as follows:

$$\begin{bmatrix} \tilde{A}' \\ \tilde{B}' e^{i\delta'} \end{bmatrix} = \frac{\cancel{e^{i\phi_{A'}}}}{\sqrt{|A'|^2 + |B'|^2}} \begin{bmatrix} |A'| \\ |B'| e^{i\delta'} \end{bmatrix}$$

This is the Jones vector that is consistent with E'_{eff}. The uninteresting overall phase factor $e^{i\phi_{A'}}$ can be incorporated into E'_{eff}, making $\tilde{A}'$ real and positive. δ' is the phase difference between B' and A'.

6.5 Jones Matrix for a Polarizer

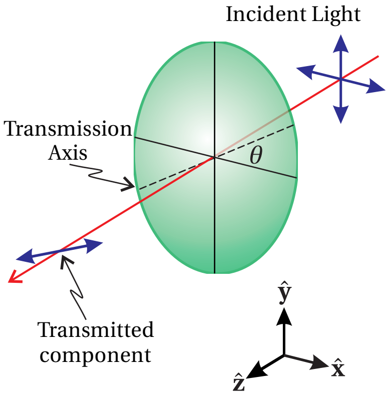


Figure 6.5 Light transmitting through a polarizer oriented with transmission axis at angle θ from x-axis.

In this section we develop a Jones matrix for describing an ideal polarizer with its transmission axis at an arbitrary angle θ from the x-axis. We will do this in a general context so that we can take advantage of the present work when discussing wave plates in the next section. To help keep things on a more conceptual level, we revert back to using electric field components directly. We will make the connection with Jones calculus at the end.

The polarizer acts on a plane wave with arbitrary polarization. The electric field of our plane wave may be written as

$$\mathbf{E}(z,t) = \left(E_x\hat{\mathbf{x}} + E_y\hat{\mathbf{y}}\right)e^{i(kz-\omega t)} \tag{6.24}$$

Let the transmission axis of the polarizer be specified by the unit vector $\hat{\mathbf{e}}_1$ and the absorption axis of the polarizer be specified by $\hat{\mathbf{e}}_2$ (orthogonal to the transmission axis). The vector $\hat{\mathbf{e}}_1$ is oriented at an angle θ from the x-axis, as shown in Fig. 6.6. We need to write the electric field components in terms of the new basis specified by $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$. By inspection of the geometry, the x-y unit vectors are connected to the new coordinate system via:

$$\begin{aligned}\hat{\mathbf{x}} &= \cos\theta\hat{\mathbf{e}}_1 - \sin\theta\hat{\mathbf{e}}_2\\ \hat{\mathbf{y}} &= \sin\theta\hat{\mathbf{e}}_1 + \cos\theta\hat{\mathbf{e}}_2\end{aligned} \tag{6.25}$$

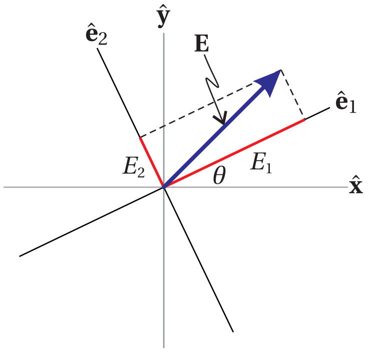


Figure 6.6 Electric field components written in the $\hat{\mathbf{e}}_1$–$\hat{\mathbf{e}}_2$ basis.

Substitution of (6.25) into (6.24) yields for the electric field

$$\mathbf{E}(z,t) = \left(E_1\hat{\mathbf{e}}_1 + E_2\hat{\mathbf{e}}_2\right)e^{i(kz-\omega t)} \tag{6.26}$$

where

$$\begin{aligned}E_1 &\equiv E_x\cos\theta + E_y\sin\theta\\ E_2 &\equiv -E_x\sin\theta + E_y\cos\theta\end{aligned} \tag{6.27}$$

Now we introduce the effect of the polarizer on the field: E_1 is transmitted unaffected, while E_2 is extinguished. To account for the effect of the device, we multiply E_2 by a parameter ξ. In the case of the polarizer, ξ is zero, but when we consider wave plates we will use other values for ξ. After traversing the polarizer, the field becomes

$$\mathbf{E}_{\text{after}}(z,t) = \left(E_1\hat{\mathbf{e}}_1 + \xi E_2\hat{\mathbf{e}}_2\right)e^{i(kz-\omega t)} \tag{6.28}$$

We now have the field after the polarizer, but it would be nice to rewrite it in terms of the original x–y basis. By inverting (6.25), or again by inspection of Fig. 6.6, we see that

$$\begin{aligned}\hat{\mathbf{e}}_1 &= \cos\theta\hat{\mathbf{x}} + \sin\theta\hat{\mathbf{y}}\\ \hat{\mathbf{e}}_2 &= -\sin\theta\hat{\mathbf{x}} + \cos\theta\hat{\mathbf{y}}\end{aligned} \tag{6.29}$$

Substitution of these relationships into (6.28) together with the definitions (6.27) for E_1 and E_2 yields

$$\begin{aligned}\mathbf{E}_{\text{after}}(z,t) &= \left[\left(E_x\cos\theta + E_y\sin\theta\right)\left(\cos\theta\hat{\mathbf{x}}+\sin\theta\hat{\mathbf{y}}\right)\right.\\ &\qquad \left.+\xi\left(-E_x\sin\theta+E_y\cos\theta\right)\left(-\sin\theta\hat{\mathbf{x}}+\cos\theta\hat{\mathbf{y}}\right)\right]e^{i(kz-\omega t)}\\ &= \left[E_x\left(\cos^2\theta+\xi\sin^2\theta\right)+E_y\left(\sin\theta\cos\theta-\xi\sin\theta\cos\theta\right)\right]\hat{\mathbf{x}}e^{i(kz-\omega t)}\\ &\quad +\left[E_x\left(\sin\theta\cos\theta-\xi\sin\theta\cos\theta\right)+E_y\left(\sin^2\theta+\xi\cos^2\theta\right)\right]\hat{\mathbf{y}}e^{i(kz-\omega t)}\end{aligned} \tag{6.30}$$

Notice that if $\xi = 1$ (i.e. no polarizer), then we get back exactly what we started with (i.e. (6.30) reduces to (6.24)).

To get to the Jones matrix for the polarizer, we note that (6.30) is a linear mixture of E_x and E_y which can be represented with matrix algebra. If we represent the electric field as a two-dimensional column vector with its x component in the top and its y component in the bottom (like a Jones vector), then we can rewrite (6.30) as

$$\mathbf{E}_{\text{after}}(z,t) = \begin{bmatrix}\cos^2\theta+\xi\sin^2\theta & \sin\theta\cos\theta-\xi\sin\theta\cos\theta\\ \sin\theta\cos\theta-\xi\sin\theta\cos\theta & \sin^2\theta+\xi\cos^2\theta\end{bmatrix}\begin{bmatrix}E_x\\ E_y\end{bmatrix}e^{i(kz-\omega t)} \tag{6.31}$$

The matrix here is a properly normalized Jones matrix, even though we did not bother factoring out E_{eff} to make a properly normalized Jones vector, as specified in (6.5). We can now write down the Jones matrix for a polarizer by inserting $\xi = 0$ into the matrix:

$$\begin{bmatrix}\cos^2\theta & \sin\theta\cos\theta\\ \sin\theta\cos\theta & \sin^2\theta\end{bmatrix} \quad \text{(polarizer with transmission axis at angle } \theta\text{)} \tag{6.32}$$

Notice that when $\theta = 0$ this matrix reduces to that of a horizontal polarizer (6.19), and when $\theta = \pi/2$, it reduces to that of a vertical polarizer (6.20).

Edwin H. Land (1909–1991, American) was born in Bridgeport, Connecticut. He began college at Harvard University, but dropped out to work on his idea of making an inexpensive polarizer. He had access to scientific literature at New York Public Library. He gained access to laboratory equipment by sneaking into Columbia University after hours. In a major breakthrough, Land invented what later would be called polaroid film. He resumed his studies at Harvard, but never graduated. This was in spite of the efforts of his wife who would extract answers from him and write up his homework. A few years later, Land and a financial backer formed Polaroid Corporation, which had tremendous success and growth thanks to Land's continued innovations over the years, including his development of an instant camera. Land would often work on a problem for days without going home or changing his clothes. He sometimes needed to be reminded to eat. (Wikipedia)

6.6 Jones Matrix for Wave Plates

We next consider *wave plates* (or *retarders*), which are usually made from birefringent crystals. The index of refraction in the crystal depends on the orientation of the electric field polarization. A wave plate has the appearance of a thin window through which the light passes. The crystal is cut such that the wave plate has a *fast* and a *slow axis*, which are 90° apart in the plane of the window. If the light is polarized along the fast axis, it experiences index n_{fast}. The orthogonal polarization component experiences higher index n_{slow}.

When a plane wave passes through a wave plate, the component of the electric field oriented along the fast axis travels faster than its orthogonal counterpart, which introduces a relative phase between the two polarization components. As light passes through a wave plate of thickness d, the phase difference that accumulates between the fast and the slow polarization components is

$$k_{\text{slow}}d - k_{\text{fast}}d = \frac{2\pi d}{\lambda_{\text{vac}}}\left(n_{\text{slow}}-n_{\text{fast}}\right) \tag{6.33}$$

By adjusting the thickness of the wave plate, one can introduce any desired phase difference.

The most common types of wave plates are the *quarter-wave plate* and the *half wave plate*. The quarter-wave plate introduces a phase difference of

$$k_{\text{slow}}d - k_{\text{fast}}d = \pi/2 + 2\pi m \qquad \text{(quarter-wave plate)} \quad (6.34)$$

between the two polarization components, where m is an integer. This means that the polarization component along the slow axis is delayed spatially by a quarter wavelength (or five quarters, etc.).

The half wave plate introduces a phase difference of

$$k_{\text{slow}}d - k_{\text{fast}}d = \pi + 2\pi m \qquad \text{(half wave plate)} \quad (6.35)$$

where m is an integer. This means that the polarization component along the slow axis is delayed spatially by a half wavelength (or three halves, etc.). When $m = 0$ in either (6.34) or (6.35), the wave plate is said to be *zero order*.

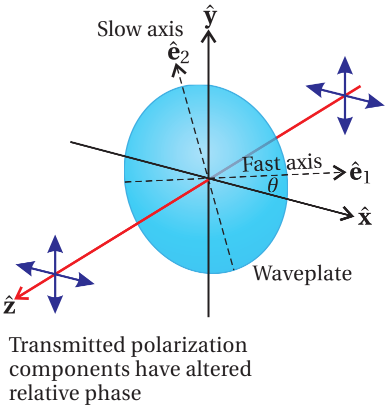


Figure 6.7 Wave plate interacting with a plane wave.

The derivation of the Jones matrix for the two wave plates is essentially the same as the derivation for the polarizer in the previous section. Let $\hat{\mathbf{e}}_1$ correspond to the fast axis, and let $\hat{\mathbf{e}}_2$ correspond to the slow axis, as illustrated in Fig. 6.7. We proceed as before. However, instead of setting ξ equal to zero in (6.31), we must choose values for ξ appropriate for each wave plate. Since nothing is absorbed, ξ should have a magnitude equal to one. The important feature is the phase of ξ. As seen in (6.33), the field component along the slow axis accumulates excess phase relative to the component along the fast axis, and we let ξ account for this. In the case of the quarter-wave plate, the appropriate factor from (6.34) is

$$\xi = e^{i\pi/2} = i \qquad \text{(quarter-wave plate)} \quad (6.36)$$

This describes a *relative* phase delay for the light emerging with polarization along the slow axis. Substituting (6.36) into (6.30) yields the Jones matrix for a quarter-wave plate:

$$\begin{bmatrix} \cos^2\theta + i\sin^2\theta & \sin\theta\cos\theta - i\sin\theta\cos\theta \\ \sin\theta\cos\theta - i\sin\theta\cos\theta & \sin^2\theta + i\cos^2\theta \end{bmatrix} \qquad \text{(quarter-wave plate)} \quad (6.37)$$

For the half-wave plate, the appropriate factor applied to the slow axis is

$$\xi = e^{i\pi} = -1 \qquad \text{(half-wave plate)} \quad (6.38)$$

and the Jones matrix becomes:

$$\begin{bmatrix} \cos^2\theta - \sin^2\theta & 2\sin\theta\cos\theta \\ 2\sin\theta\cos\theta & \sin^2\theta - \cos^2\theta \end{bmatrix} = \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix} \qquad \text{(half-wave plate)} \quad (6.39)$$

Remember that θ refers to the angle that the fast axis makes with respect to the x-axis.

Before moving on, consider the following two examples that illustrate how wave plates are often used:

Example 6.2

Calculate the Jones matrix for a quarter-wave plate at $\theta = 45°$, and determine its effect on horizontally polarized light.

Solution: At $\theta = 45°$, the Jones matrix for the quarter-wave plate (6.37) reduces to

$$\frac{e^{i\pi/4}}{\sqrt{2}}\begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \qquad \text{(quarter-wave plate, fast axis at } \theta = 45° \text{)} \tag{6.40}$$

The overall phase factor $e^{i\pi/4}$ in front is unimportant since it can be adjusted arbitrarily by moving the light source forwards or backwards through a fraction of a wavelength.

Now we calculate the effect of the quarter-wave plate (oriented at $\theta = 45°$) operating on horizontally polarized light:

$$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -i \end{bmatrix} \tag{6.41}$$

The linearly polarized light becomes right-circularly polarized (see Table 6.1)

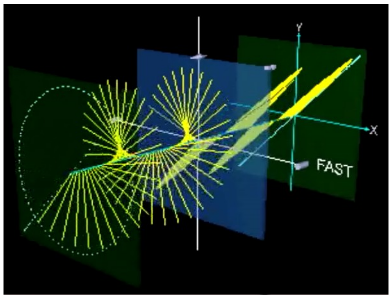


Figure 6.8 Animation showing effects of polarizers and wave plates on polarized light.

The previous example shows that a quarter-wave plate (properly oriented) can change linearly polarized light into circularly polarized light. A quarter-wave plate can perform the reverse operation as well. On the other hand, as seen in the next example, a half-wave plate can rotate the polarization angle of linearly polarized light by varying degrees while preserving the linear polarization.

Example 6.3

Calculate the effect of a half-wave plate at an arbitrary θ on horizontally polarized light.

Solution: Multiplying by the half-wave matrix (6.39), we obtain

$$\begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \cos 2\theta \\ \sin 2\theta \end{bmatrix} \tag{6.42}$$

The resulting Jones vector describes linearly polarized light at an angle $\alpha = 2\theta$ from the x-axis (see Table 6.1).

6.7 Polarization Effects of Reflection and Transmission

When light encounters a material interface, the amount of reflected and transmitted light depends on the polarization. The Fresnel coefficients (3.20)–(3.23) dictate how much of each polarization is reflected and how much is transmitted. In addition, the Fresnel coefficients keep track of phases intrinsic in the reflection phenomenon. This is true also for reflections from multilayer coatings with effective Fresnel coefficients (4.59), (4.60), (4.64) and (4.65).

To the extent that the s and p components of the field behave differently, the overall polarization state is altered. For example, a linearly-polarized field upon reflection can become elliptically polarized (see L 6.9). Even when a wave reflects at normal incidence so that the s and p components are indistinguishable, right-circular polarized light becomes left-circular polarized. This is the same effect that causes a right-handed person to appear left-handed when viewed in a mirror.

We can use Jones calculus to keep track of how reflection and transmission influences polarization. However, before proceeding, we emphasize that in this context we do not strictly adhere to a single coordinate system as we did in chapter 3, for example in Fig. 3.1. Instead, we consider each plane wave, whether incident, reflected or transmitted, to propagate in the z-direction of its own frame, regardless of the relative angles between the incident and reflected wave. This loose manner of defining coordinate systems, depicted in Fig. 6.9, has a great advantage. The x and y dimensions in each individual frame are aligned parallel to their respective s and p field component. We will adopt the convention that p-polarized light in all cases is associated with the x-dimension (horizontal, say). The s-polarized component then lies along the y-dimension (vertical). These conventions are different from those used in chapter 3 but will do us no harm.

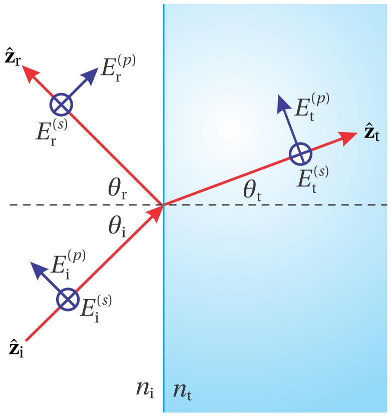


Figure 6.9 Incident, reflected and transmitted plane waves, each propagating along the z-axis of its own reference frame.

We are now in a position to see why there is a handedness inversion upon reflection from a mirror. Notice in Fig. 6.9 that for the incident light, the s-component of the field crossed into the p-component of the field yields a vector pointing along the beam's propagation direction. However, for the reflected light, the s-component crossed into the p-component points opposite to that beam's propagation direction.

The Jones matrix corresponding to reflection from a surface is

$$\begin{bmatrix} -r_p & 0 \\ 0 & r_s \end{bmatrix} \qquad \text{(Jones matrix for reflection)} \quad (6.43)$$

By convention, we place the minus sign on the coefficient r_p to take care of handedness inversion. We could put the minus sign on r_s instead; the important point is that the two polarizations acquire a relative phase differential of π when the propagation direction flips.[3]

The Fresnel coefficients specify the ratios of the exiting fields to the incident ones. When (6.43) operates on an arbitrary Jones vector such as (6.11), $-r_p$ multiplies the horizontal component of the field, and r_s multiplies the vertical component of the field. Especially in the case of reflection from an absorbing surface such as a metal, the phases of the two polarization components can vary markedly (see P6.13). Thus, linearly polarized light containing both s- and p-components in general becomes elliptically polarized when reflected from such a surface. When light undergoes total internal reflection, again the phases of the s-

---

[3] The minus sign is needed for our specific convention of field directions, as drawn in Fig. 6.9. In our convention, r_s and r_p are identical at normal incidence.

and p-components differ markedly, which can cause linearly polarized light to become elliptically polarized (see P6.14).

Transmission through a material interface can also influence the polarization of the field, although typically to a lesser degree. However, there is no handedness inversion, since the light continues on in a forward sense. The Jones matrix for transmission is

$$\begin{bmatrix} t_p & 0 \\ 0 & t_s \end{bmatrix} \qquad \text{(Jones matrix for transmission)} \tag{6.44}$$

If a beam of light encounters a series of mirrors, the final polarization is determined by multiplying the sequence of appropriate Jones matrices (6.43) onto the initial polarization. This procedure is straightforward if the normals to all of the mirrors lie in a single plane (say parallel to the surface of an optical bench). However, if the beam path deviates from this plane (due to vertical tilt on the mirrors), then we must reorient our coordinate system before each mirror to have a new 'horizontal' (p-polarized dimension) and the new 'vertical' (s-polarized dimension). Earlier in this chapter we performed a rotation of a coordinate system through an angle θ, described in (6.27), which is also useful here. The rotation can be accomplished by multiplying the following matrix onto the incident Jones vector:

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \qquad \text{(rotation of coordinates through an angle } \theta) \tag{6.45}$$

This is understood as a rotation about the z-axis. The angle of rotation θ is chosen such that the rotated x-axis lies in the plane of incidence for the mirror. When such a reorientation of coordinates is necessary, the two orthogonal field components in the initial coordinate system are stirred together to form the field components in the new system. This does not change the intrinsic characteristics of the polarization, just its representation.

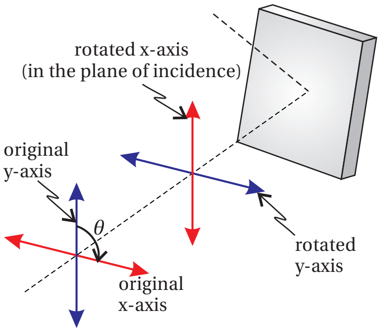


Figure 6.10 If the plane of incidence does not coincide for successive elements in an optical system, a rotation matrix must be applied to rotate the x-axis to the plane of incidence before computing the effect of each element.

Appendix 6.A Ellipsometry

Measuring the polarization of light reflected from a surface can yield information regarding the optical properties of that surface. As done in L 6.9, it is possible to characterize the polarization of a beam of light using a quarter-wave plate and a polarizer. However, we often want to measure reflections at a range of frequencies, and this would require a different quarter-wave plate thickness d for each wavelength used (see (6.34)). Therefore, many commercial *ellipsometers* do not try to extract the helicity of the light, but only the ellipticity. In this case only polarizers are needed, which can be made to work over a wide range of wavelengths. If, in addition, a variety of incident angles are measured, it is possible to extract detailed information about the optical constants n and κ and the thicknesses of possibly many layers of materials influencing the reflection.

Commercial ellipsometers[4] typically employ two polarizers, one before and one after the sample, where s- and p-polarized reflections take place. The first polarizer ensures that linearly polarized light arrives at the test surface (polarized at angle α to give both s- and p-components). The Jones matrix for the test surface reflection is given by (6.43), and the Jones matrix for the analyzing polarizer oriented at angle θ is given by (6.32). The Jones vector for the light arriving at the detector is then

---

[4] See Spectroscopic Ellipsometry Tutorial at J. A. Woollam Co.

$$\begin{aligned}\begin{bmatrix}\cos^2\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & \sin^2\theta\end{bmatrix}\begin{bmatrix}-r_p & 0 \\ 0 & r_s\end{bmatrix}\begin{bmatrix}\cos\alpha \\ \sin\alpha\end{bmatrix} \\ = \begin{bmatrix}-r_p\cos\alpha\cos^2\theta + r_s\sin\alpha\sin\theta\cos\theta \\ -r_p\cos\alpha\sin\theta\cos\theta + r_s\sin\alpha\sin^2\theta\end{bmatrix}\end{aligned} \tag{6.46}$$

and the intensity arriving to the detector is

$$\begin{aligned}I &\propto \left|-r_p\cos\alpha\cos^2\theta + r_s\sin\alpha\cos\theta\sin\theta\right|^2 + \left|-r_p\cos\alpha\cos\theta\sin\theta + r_s\sin\alpha\sin^2\theta\right|^2 \\ &= \left|r_p\right|^2\cos^2\alpha\cos^2\theta + |r_s|^2\sin^2\alpha\sin^2\theta - \frac{\left(r_p r_s^* + r_s r_p^*\right)}{4}\sin 2\alpha\sin 2\theta\end{aligned} \tag{6.47}$$

For ellipsometry measurements, it is customary to express the ratio of Fresnel coefficients as

$$r_p/r_s \equiv \tan\Psi e^{i\Delta} \tag{6.48}$$

In this case, the intensity may be shown to be proportional to (see problem P6.15)

$$I \propto 1 - \eta\sin 2\theta + \xi\cos 2\theta \tag{6.49}$$

where

$$\eta \equiv 2\frac{\tan\Psi\cos\Delta\tan\alpha}{\tan^2\Psi + \tan^2\alpha} \quad \text{and} \quad \xi \equiv \frac{\tan^2\Psi - \tan^2\alpha}{\tan^2\Psi + \tan^2\alpha} \tag{6.50}$$

In commercial ellipsometers, the angle θ of the analyzing polarizer often rotates at a high speed, and the time dependence of the light reaching a detector is analyzed. From this type of measurement, the coefficients η and ξ can be extracted with high precision. Then equations (6.50) can be inverted (see problem P6.15) to reveal

$$\tan\Psi = \sqrt{\frac{1+\xi}{1-\xi}}\,|\tan\alpha| \quad \text{and} \quad \cos\Delta = \frac{\eta}{\sqrt{1-\xi^2}}\operatorname{sign}(\alpha) \tag{6.51}$$

From a series of these types of measurements, it is possible to extract the values of n and κ for materials from the expressions for r_s and r_p (with the aid of a computer!). With a sufficiently large number of unique measurements, it is possible even to characterize multilayer coatings involving layers with varying thicknesses and indices.

Appendix 6.B Partially Polarized Light

We outline here an approach for dealing with partially polarized light, which is a mixture of *polarized* and *unpolarized* light. Most natural light such as sunshine is unpolarized. The transverse electric field direction in natural light varies rapidly (and quasi randomly). Such variations imply the superposition of multiple frequencies as opposed to the single frequency assumed in the formulation of Jones calculus earlier in this chapter. Unpolarized light can become partially polarized when it, for example, reflects from a surface at oblique incidence, since s and p components of the polarization might reflect with differing strength.

Stokes vectors are used to keep track of the partial polarization (and attenuation) of a light beam as the light progresses through an optical system.[5] In contrast, Jones vectors are designed for pure polarization states. We can consider any light beam as an intensity sum of completely unpolarized light and perfectly polarized light:

$$I = I_{\text{pol}} + I_{\text{un}} \tag{6.52}$$

It is assumed that both types of light propagate in the same direction.

The main characteristic of unpolarized light is that it cannot be extinguished by a single polarizer (even in combination with a wave plate). Moreover, the transmission of unpolarized light through an ideal polarizer is always 50%. On the other hand, polarized light (be it linearly, circularly, or elliptically polarized) can always be represented by a Jones vector, and it is always possible to extinguish it using a quarter-wave plate and a single polarizer.

We may introduce the *degree of polarization* as the fraction of the intensity that is in a definite polarization state:

$$\xi\text{pol} \equiv \frac{I_{\text{pol}}}{I_{\text{pol}} + I_{\text{un}}} \tag{6.53}$$

The degree of polarization takes on values between zero and one. Thus, if the light is completely unpolarized (such that $I_{\text{pol}} = 0$), the degree of polarization is zero, and if the beam is fully polarized (such that $I_{\text{un}} = 0$), the degree of polarization is one.

A Stokes vector, which characterizes a partially polarized beam, is written as

$$\begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix}$$

The parameter

$$S_0 \equiv \frac{I}{I_{\text{in}}} = \frac{I_{\text{pol}} + I_{\text{un}}}{I_{\text{in}}} \tag{6.54}$$

is a comparison of the beam's intensity (or power) to a benchmark or 'input' intensity, I_{in}, measured before the beam enters the optical system under consideration. I represents the intensity at the point of investigation, where one wishes to characterize the beam. Thus, the value $S_0 = 1$ represents the input intensity,

Sir George Gabriel Stokes (1819–1903, Irish) was born in Skreen, Ireland. He entered Cambridge University at age 18 and graduated four years later with the distinction of senior wrangler. In 1849, he became a professor of mathematics at Cambridge where he later worked with James Clerk Maxwell and Lord Kelvin to form the Cambridge School of Mathematical Physics. Stokes was a powerful mathematician as well as a good experimentalist, often testing his theoretical solutions in the laboratory. In addition to his contributions to optics, Stokes made important contributions to fluid dynamics (e.g. the Navier-Stokes equations) and to mathematical physics; Stokes' theorem is employed several places in this in this book.(Wikipedia) and S_0 can drop to values less than one, to account for attenuation of light by polarizers in the system. (S_0 could increase in the atypical case of amplification.)

---

[5] E. Hecht, *Optics*, 3rd ed., Sect. 8.12.1 (Massachusetts: Addison-Wesley, 1998).

The next parameter, S_1, describes how much the light looks either horizontally or vertically polarized, and it is defined as

$$S_1 \equiv \frac{2I_{\text{hor}}}{I_{\text{in}}} - S_0 \tag{6.55}$$

Here, I_{hor} represents the amount of light detected if an ideal linear polarizer is placed with its axis aligned horizontally directly in front of the detector (inserted where the light is characterized). S_1 ranges between negative one and one, taking on its extremes when the light is linearly polarized either horizontally or vertically, respectively. If the light has been attenuated, it may still be perfectly horizontally polarized even if S_1 has a magnitude less than one. (Alternatively, one might examine S_1/S_0, which is guaranteed to range between negative one and one.)

The parameter S_2 describes how much the light looks linearly polarized along the diagonals. It is given by

$$S_2 \equiv \frac{2I_{45^\circ}}{I_{\text{in}}} - S_0 \tag{6.56}$$

Similar to the previous case, I_{45° represents the amount of light detected if an ideal linear polarizer is placed with its axis at 45° directly in front of the detector (inserted where the light is characterized). As before, S_2 ranges between negative one and one, taking on extremes when the light is linearly polarized either at 45° or 135°.

Finally, S_3 characterizes the extent to which the beam is either right or left circularly polarized:

$$S_3 \equiv \frac{2I_{\text{r-cir}}}{I_{\text{in}}} - S_0 \tag{6.57}$$

Here, $I_{\text{r-cir}}$ represents the amount of light detected if an ideal right-circular polarizer is placed directly in front of the detector. A right-circular polarizer is one that passes right-handed polarized light, but blocks left handed polarized light. One way to construct such a polarizer is a quarter-wave plate, followed by a linear polarizer with the transmission axis aligned 45° from the wave-plate fast axis, followed by another quarter-wave plate at −45° from the polarizer (see P6.12).[6] Again, this parameter ranges between negative one and one, taking on the extremes for right and left circular polarization, respectively.

Importantly, if any of the parameters S_1, S_2, or S_3 take on their extreme values (i.e. a magnitude equal to S_0), the other two parameters necessarily equal zero. As an example, if a beam is linearly polarized in the horizontal direction with $I = I_{\text{in}}$, then we have $I_{\text{hor}} = I_{\text{in}}$, $I_{45^\circ} = I_{\text{in}}/2$, and $I_{\text{r-cir}} = I_{\text{in}}/2$. This yields $S_0 = 1$, $S_1 = 1$, $S_2 = 0$, and $S_3 = 0$. As a second example, suppose that the light has been attenuated to $I = I_{\text{in}}/3$ but is purely left circularly polarized. Then we have

---

[6] The final quarter-wave plate is to put the light back into the original circular state – not needed to measure the Stokes parameter.

$I_{\text{hor}} = I_{\text{in}}/6$, $I_{45^\circ} = I_{\text{in}}/6$, and $I_{\text{r-cir}} = 0$. Whereas the Stokes parameters are $S_0 = 1/3$, $S_1 = 0$, $S_2 = 0$, and $S_3 = -1/3$.

Another interesting case is completely unpolarized light, which transmits 50% through all of the polarizers discussed above. In this case, $I_{\text{hor}} = I_{45^\circ} = I_{\text{r-cir}} = I/2$ and $S_1 = S_2 = S_3 = 0$.

Example 6.4

Find the Stokes parameters for perfectly polarized light, represented by an arbitrary Jones vector $\left[\begin{smallmatrix} A \\ B \end{smallmatrix}\right]$ where A and B are complex.[7] Depending on the values A and B, the polarization can follow any ellipse.

Solution: The input intensity of this polarized beam is $I_{\text{in}} = I_{\text{pol}} = |A|^2 + |B|^2$, according to Eq. (6.23), where we absorb the factor $\frac{1}{2}\epsilon_0 c |E_{\text{eff}}|^2$ into $|A|^2$ and $|B|^2$ for convenience. The Jones vector for the light that passes through a horizontal polarizer is

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} A \\ 0 \end{bmatrix}$$

which gives a measured intensity of $I_{\text{hor}} = |A|^2$. Similarly, the Jones vector when the beam is passed through a polarizer oriented at 45° is

$$\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \frac{A+B}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

leading to an intensity of

$$I_{45^\circ} = \frac{|A+B|^2}{2} = \frac{|A|^2 + |B|^2 + A^*B + AB^*}{2}$$

Finally, the Jones vector for light passing through a right-circular polarizer (see P6.12) is

$$\frac{1}{2}\begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \frac{A+iB}{2}\begin{bmatrix} 1 \\ -i \end{bmatrix}$$

giving an intensity of

$$I_{\text{r-cir}} = \frac{|A+iB|^2}{2} = \frac{|A|^2 + |B|^2 + i\,(A^*B - AB^*)}{2}$$

Thus, the Stokes parameters become

$$S_0 = \frac{|A|^2 + |B|^2}{I_{\text{in}}} = 1$$

$$S_1 = \frac{2|A|^2}{I_{\text{in}}} - \frac{|A|^2 + |B|^2}{I_{\text{in}}} = \frac{|A|^2 - |B|^2}{I_{\text{in}}}$$

$$S_2 = \frac{|A|^2 + |B|^2 + A^*B + AB^*}{I_{\text{in}}} - \frac{|A|^2 + |B|^2}{I_{\text{in}}} = \frac{A^*B + AB^*}{I_{\text{in}}}$$

$$S_3 = \frac{|A|^2 + |B|^2 + i\,(A^*B - AB^*)}{I_{\text{in}}} - \frac{|A|^2 + |B|^2}{I_{\text{in}}} = i\frac{(A^*B - AB^*)}{I_{\text{in}}}$$

---

[7] We will find it easier in this appendix to write $\begin{bmatrix} A \\ B \end{bmatrix}$ instead of $\begin{bmatrix} |A| \\ |B|e^{i\delta} \end{bmatrix}$, where δ is the phase difference between B and A.

Note that the unpolarized portion of the light does not contribute to S_1, S_2, or S_3. Half of the unpolarized light survives any of the test filters, which cancels neatly with the unpolarized portion of $S_0 = \frac{I_{\text{pol}}+I_{\text{un}}}{I_{\text{in}}}$ in Eqs. (6.55)–(6.57).

With the aid of the results in Example 6.4, a completely general form of the Stokes vector may then be written as

$$\begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix} = \frac{1}{I_{\text{in}}} \begin{bmatrix} I_{\text{pol}} + I_{\text{un}} \\ |A|^2 - |B|^2 \\ A^* B + AB^* \\ i\left(A^* B - AB^*\right) \end{bmatrix} \tag{6.58}$$

where the Jones vector for the polarized portion of the light is

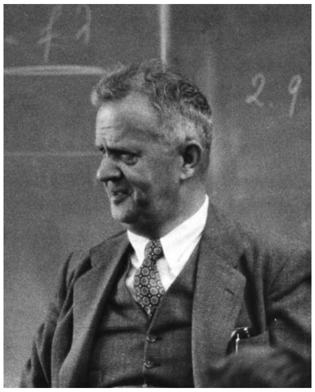
$$\begin{bmatrix} A \\ B \end{bmatrix}$$

and the intensity of the polarized portion of the light is

$$I_{\text{pol}} = |A|^2 + |B|^2 \tag{6.59}$$

Again, we have hidden the factor $\frac{1}{2}\epsilon_0 c\,|E_{\text{eff}}|^2$ for the polarized portion of the light inside $|A|^2$ and $|B|^2$.

We would like to express the degree of polarization in terms of the Stokes parameters. We first note that the quantity $\sqrt{S_1^2 + S_2^2 + S_3^2}$ can be expressed as



Hans Mueller (1900-1965, Swiss) was a shepherd until his late teens. As a physics professor at MIT, he built on the work of Stokes and in 1943 formulated a matrix method for manipulating Stokes vectors. He was an engaging lecturer into the 1950s and was known for his exciting demonstrations. He was a student of Arnold Sommerfeld, and did seminal work on ferroelectricity (he is reported to have coined the term). See Laszlo Tisza,"Adventures of a Theoretical Physicist, Part II: America," Phys. Perspect. **11**, 120-168 (2009).

$$\begin{aligned} \sqrt{S_1^2 + S_2^2 + S_3^2} &= \sqrt{\left(\frac{|A|^2 - |B|^2}{I_{\text{in}}}\right)^2 + \left(\frac{\left(A^* B + AB^*\right)}{I_{\text{in}}}\right)^2 + \left(\frac{i\left(A^* B - AB^*\right)}{I_{\text{in}}}\right)^2} \\ &= \frac{|A|^2 + |B|^2}{I_{\text{in}}} \\ &= \frac{I_{\text{pol}}}{I_{\text{in}}} \end{aligned} \tag{6.60}$$

Substituting (6.54) and (6.60) into the expression for the degree of polarization (6.53) yields

$$\xi\text{pol} \equiv \frac{1}{S_0}\sqrt{S_1^2 + S_2^2 + S_3^2} \tag{6.61}$$

If the light is polarized such that it perfectly transmits through or is perfectly extinguished by one of the three test polarizers associated with S_1, S_2, or S_3, then the degree of polarization will be unity. Obviously, it is possible to have pure polarization states that are not aligned with the axes of any one of these test polarizers. In this situation, the degree of polarization is still one, although the values S_1, S_2, and S_3 may all three contribute to (6.61).

Finally, it is possible to represent polarizing devices as matrices that operate on the Stokes vectors in much the same way that Jones matrices operate on Jones vectors. Since Stokes vectors are four-dimensional, the matrices used are four-by-four. These are known as *Mueller matrices*.[8]

---

[8] E. Hecht, *Optics*, 3rd ed., Sect. 8.12.3 (Massachusetts: Addison-Wesley, 1998).

Derivation: Mueller Matrix for a Linear Polarizer

We know that 50% of the unpolarized light transmits through a polarizer, ending up as polarized light with Jones vector

$$\begin{bmatrix} A_1' \\ B_1' \end{bmatrix} = \sqrt{\frac{I_{\text{un}}}{2}} \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$$

(see table 6.1). As usual, let θ give the angle of the transmission axis relative to the horizontal. The Jones matrix (6.23) acts on the polarized portion of the light as follows

$$\begin{bmatrix} A_2' \\ B_2' \end{bmatrix} = \begin{bmatrix} \cos^2\theta & \cos\theta\sin\theta \\ \cos\theta\sin\theta & \sin^2\theta \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = [A\cos\theta + B\sin\theta] \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$$

One might be tempted to add $\left[\begin{smallmatrix} A_1' \\ B_1' \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} A_2' \\ B_2' \end{smallmatrix}\right]$, but this would be wrong, since the two beams are not coherent. As mentioned previously, unpolarized light necessarily contains multiple frequencies, and so the fields from the polarized and unpolarized beams destructively interfere as often as they constructively interfere. In this case, we simply add intensities rather than fields. That is, we have

$$\begin{aligned} \left|A'\right|^2 = \left|A_1'\right|^2 + \left|A_2'\right|^2 &= \left[\frac{I_{\text{un}}}{2} + |A\cos\theta + B\sin\theta|^2\right]\cos^2\theta \\ &= \left[\frac{I_{\text{un}}}{2} + |A|^2\cos^2\theta + |B|^2\sin^2\theta + \left(A^*B + AB^*\right)\sin\theta\cos\theta\right]\cos^2\theta \\ &= I_{\text{in}}\left[\frac{S_0}{2} + \frac{\cos 2\theta}{2}S_1 + \frac{\sin 2\theta}{2}S_2\right]\cos^2\theta \end{aligned}$$

Similarly,

$$\left|B'\right|^2 = \left|B_1'\right|^2 + \left|B_2'\right|^2 = I_{\text{in}}\left[\frac{S_0}{2} + \frac{\cos 2\theta}{2}S_1 + \frac{\sin 2\theta}{2}S_2\right]\sin^2\theta$$

Since the light has gone through a linear polarizer, we are guaranteed that A' and B' have the same phase. Therefore, $A'^*B' = A'B'^* = |A'||B'|$. In view of (6.58), these results lead to

$$\begin{aligned} S_0' &= \frac{\left|A'\right|^2 + \left|B'\right|^2}{I_{\text{in}}} = \frac{S_0}{2} + \frac{\cos 2\theta}{2}S_1 + \frac{\sin 2\theta}{2}S_2 \\ S_1' &= \frac{\left|A'\right|^2 - \left|B'\right|^2}{I_{\text{in}}} = \left[\frac{S_0}{2} + \frac{\cos 2\theta}{2}S_1 + \frac{\sin 2\theta}{2}S_2\right]\left(\cos^2\theta - \sin^2\theta\right) \\ &= \frac{\cos 2\theta}{2}S_0 + \frac{\cos^2 2\theta}{2}S_1 + \frac{\sin 4\theta}{4}S_2 \\ S_2' &= \frac{\left|A'\right|\left|B'\right| + \left|A'\right|\left|B'\right|}{I_{\text{in}}} = 2\left[\frac{S_0}{2} + \frac{\cos 2\theta}{2}S_1 + \frac{\sin 2\theta}{2}S_2\right]\cos\theta\sin\theta \\ &= \frac{\sin 2\theta}{2}S_0 + \frac{\sin 4\theta}{4}S_1 + \frac{\sin^2 2\theta}{2}S_2 \\ S_3' &= i\frac{\left|A'\right|\left|B'\right| - \left|A'\right|\left|B'\right|}{I_{\text{in}}} = 0 \end{aligned}$$

These transformations expressed in matrix format become

$$\begin{bmatrix} S_0' \\ S_1' \\ S_2' \\ S_3' \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & \cos 2\theta & \sin 2\theta & 0 \\ \cos 2\theta & \cos^2 2\theta & \frac{1}{2}\sin 4\theta & 0 \\ \sin 2\theta & \frac{1}{2}\sin 4\theta & \sin^2 2\theta & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix}$$

which reveals the Mueller matrix for a linear polarizer.

The Mueller matrix for a half-wave plate is worked out below. The Mueller matrix for a quarter-wave plate is deferred to problem P6.16

Derivation: Mueller Matrix for a Half-Wave Plate

We know that all of the light transmits through the wave plate. This immediately gives

$$S_0' = S_0$$

The wave plate does nothing to unpolarized light. On the other hand, the polarized portion of the light is influenced by the wave plate as follows (see (6.39)):

$$\begin{bmatrix} A' \\ B' \end{bmatrix} = \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix}\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} A\cos 2\theta + B\sin 2\theta \\ A\sin 2\theta - B\cos 2\theta \end{bmatrix}$$

As usual, θ is the angle of the fast axis relative to the horizontal. (As expected, $\left|A'\right|^2 + \left|B'\right|^2 = |A|^2 + |B|^2$; the intensity of the light is unaltered.) Using (6.58) we get

$$\begin{aligned}
S_1' &= \frac{\left|A'\right|^2 - \left|B'\right|^2}{I_\text{in}} = \frac{|A\cos 2\theta + B\sin 2\theta|^2 - |A\sin 2\theta - B\cos 2\theta|^2}{I_{in}} \\
&= \frac{\left(|A|^2 - |B|^2\right)\cos 4\theta + \left(A^* B + AB^*\right)\sin 4\theta}{I_{in}} = S_1\cos 4\theta + S_2\sin 4\theta \\
S_2' &= \frac{A'^* B' + A' B'^*}{I_{in}} \\
&= \frac{\left(A^*\cos 2\theta + B^*\sin 2\theta\right)\left(A\sin 2\theta - B\cos\theta\right)}{I_{in}} \\
&+ \frac{\left(A\cos 2\theta + B\sin 2\theta\right)\left(A^*\sin 2\theta - B^*\cos\theta\right)}{I_{in}} \\
&= \frac{|A|^2 - |B|^2}{I_{in}}\sin 4\theta - \frac{AB^* + A^* B}{I_{in}}\cos 4\theta = S_1\sin 4\theta - S_2\cos 4\theta \\
S_3' &= i\frac{A'^* B' - A' B'^*}{I_{in}} \\
&= i\frac{\left(A^*\cos 2\theta + B^*\sin 2\theta\right)\left(A\sin 2\theta - B\cos\theta\right)}{I_{in}} \\
&- i\frac{\left(A\cos 2\theta + B\sin 2\theta\right)\left(A^*\sin 2\theta - B^*\cos\theta\right)}{I_{in}} \\
&= -i\frac{A^* B - AB^*}{I_{in}} = -S_3
\end{aligned}$$

These transformations expressed in matrix format become

$$\begin{bmatrix} S_0' \\ S_1' \\ S_2' \\ S_3' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos 4\theta & \sin 4\theta & 0 \\ 0 & \sin 4\theta & -\cos 4\theta & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix}$$

which reveals the Mueller matrix for a half-wave plate.

Exercises

Exercises for 6.2 Jones Vectors for Representing Polarization

P6.1 Show that $\left(A\hat{\mathbf{x}}+Be^{i\delta}\hat{\mathbf{y}}\right)\cdot\left(A\hat{\mathbf{x}}+Be^{i\delta}\hat{\mathbf{y}}\right)^{*}=1$, as defined in connection with (6.5).

P6.2 Prove that if $0<\delta<\pi$, the helicity is left-handed, and if $\pi<\delta<2\pi$ the helicity is right-handed.

HINT: Write the relevant real field associated with (6.5)

$$\mathbf{E}(z,t)=|E_{\text{eff}}|\left[\hat{\mathbf{x}}A\cos\left(kz-\omega t+\phi\right)+\hat{\mathbf{y}}B\cos\left(kz-\omega t+\phi+\delta\right)\right]$$

where ϕ is the phase of E_{eff}. Freeze time at, say, $t=\phi/\omega$. Determine the field, for example, at $z=0$ and at $z=\lambda/4$ (a quarter cycle downstream). If $\mathbf{E}(0,t)\times\mathbf{E}(\lambda/4,t)$ points in the direction of $\mathbf{k}$, then the helicity matches that of a common wood screw.

L6.3 Determine how much right-handed circularly polarized light ($\lambda_{\text{vac}}=$ 633 nm) is delayed (or advanced if ϕ is negative) with respect to left-handed circularly polarized light as it goes through approximately 3 cm of Karo syrup (the neck of the bottle). This phenomenon is called *optical activity*. Because of a definite-handedness to the molecules in the syrup, right- and left-handed polarized light experience slightly different refractive indices. (video)

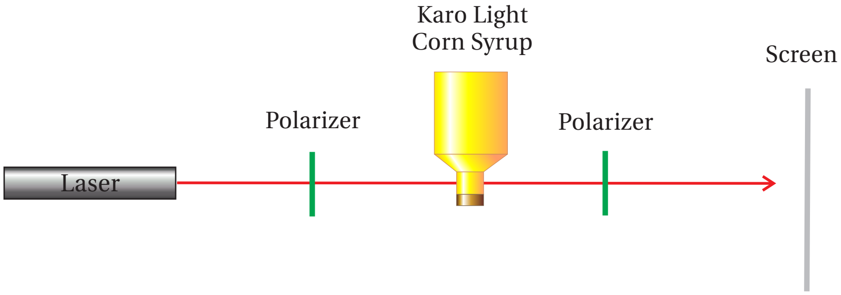


Figure 6.11 Lab schematic for L 6.3.

HINT: Linearly polarized light contains equal amounts of right and left circularly polarized light. Consider

$$\frac{1}{2}\begin{bmatrix}1\\ i\end{bmatrix}+\frac{e^{i\phi}}{2}\begin{bmatrix}1\\ -i\end{bmatrix}$$

where ϕ is the phase delay of the right circular polarization. Show that this can be written as

$$e^{i\delta}\begin{bmatrix}\cos\phi/2\\ \sin\phi/2\end{bmatrix}$$

The overall phase δ is unimportant. Compare this with

$$\begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix}$$

where α is the angle of linearly polarized light (see table 6.1).

Exercises for 6.3 Elliptically Polarized Light

P6.4 Consider the Jones vector

$$\begin{bmatrix} A \\ Be^{i\delta} \end{bmatrix}$$

For the following cases, what is the orientation of the major axis, and what is the ellipticity of the light? Case I: $A = B = 1/\sqrt{2}$; $\delta = 0$ Case II: $A = B = 1/\sqrt{2}$; $\delta = \pi/2$; Case III: $A = B = 1/\sqrt{2}$; $\delta = \pi/4$.

Exercises for 6.4 Linear Polarizers and Jones Matrices

P6.5 (a) Suppose that linearly polarized light is oriented at an angle α with respect to the horizontal or x-axis (see table 6.1). What fraction of the original *intensity* gets through a vertically oriented polarizer?

(b) If the original light is right-circularly polarized, what fraction of the original *intensity* gets through the same polarizer?

Exercises for 6.5 Jones Matrix for a Polarizer

P6.6 Horizontally polarized light ($\alpha = 0$) is sent through two polarizers, the first oriented at $\theta_1 = 45°$ and the second at $\theta_2 = 90°$.

(a) What fraction of the original intensity emerges?

(b) What is the fraction if the ordering of the polarizers is reversed?

P6.7 (a) Suppose that linearly polarized light is oriented at an angle α with respect to the horizontal or x-axis. What fraction of the original intensity emerges from a polarizer oriented with its transmission at angle θ from the x-axis?

Answer: $\cos^2(\theta - \alpha)$; compare with P6.5.

(b) If the original light is right circularly polarized, what fraction of the original intensity emerges from the same polarizer?

P6.8 Derive (6.12), (6.13), and (6.14).

HINT: Analyze the Jones vector just as you would analyze light in the laboratory. Put a polarizer in the beam and observe the intensity of the light as a function of polarizer angle. Compute the intensity via (6.23). Then find the polarizer angle (call it α) that gives a maximum (or a minimum) of intensity. The angle then corresponds to an axis of the ellipse inscribing the E-field as it spirals. When taking the arctangent, remember that it is defined only over half of the unit circle. You can add π for another valid result (which gives a second ellipse axis).

Exercises for 6.6 Jones Matrix for Wave Plates

L6.9 Create a source of unknown elliptical polarization by reflecting a linearly polarized laser beam (with both s and p-components) from a metal mirror with a large incident angle (i.e. $\theta_i \geq 80°$). Use a quarter-wave plate and a polarizer to determine the Jones vector of the reflected beam. Find the ellipticity, the helicity (right or left handed), and the orientation of the major axis. (video)

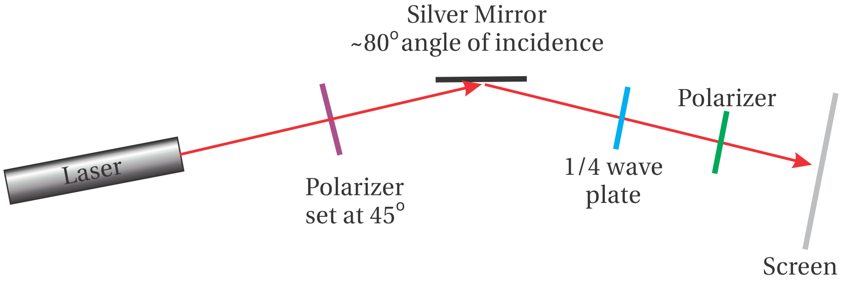


Figure 6.12 Lab schematic for L 6.9.

HINT: A polarizer alone can reveal the direction of the major and minor axes and the ellipticity, but it does not reveal the helicity. Use a quarter-wave plate (oriented at a special angle θ) to convert the unknown elliptically polarized light into linearly polarized light. A subsequent polarizer can then extinguish the light, from which you can determine the Jones vector of the light coming through the wave plate. This must equal the original (unknown) Jones vector (6.11) operated on by the wave plate (6.37). As you solve the matrix equation, it is helpful to note that the inverse of (6.37) is its own complex conjugate.

P6.10 What is the minimum thickness (called zero-order thickness) of a quartz plate made to operate as a quarter-wave plate for $\lambda_{\text{vac}} = 500$ nm? The indices of refraction are $n_{\text{fast}} = 1.54424$ and $n_{\text{slow}} = 1.55335$.

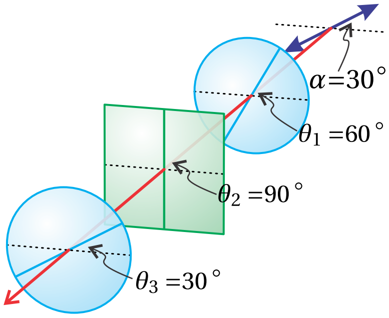


Figure 6.13 Arrangement for P6.11.

P6.11 Light that is linearly polarized along $\alpha = 30°$ traverses a quarter-wave plate with fast axis at $\theta_1 = 60°$. The light then goes through a polarizer with transmission axis at $\theta_2 = 90°$ followed by a half-wave plate with fast axis at $\theta_3 = 30°$.

(a) What is the Jones vector of the light emerging from the final element?

(b) What fraction of the original intensity transmits through the system?

P6.12 A right-circular polarizer can be constructed using a quarter-wave plate with fast axis at 45°, followed by a linear polarizer oriented vertically, and finally a quarter-wave plate with fast axis at −45°.

(a) Calculate the Jones matrix for this system.

Answer: $\frac{1}{2}\begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix}$

(b) Check that the device leaves right-circularly polarized light unaltered while killing left-circularly polarized light.

Exercises for 6.7 Polarization Effects of Reflection and Transmission

P6.13 Light is linearly polarized at $\alpha = 45°$ with a Jones vector according to table 6.1. The light is reflected from a vertical silver mirror with angle of incidence $\theta_\mathrm{i} = 80°$, as described in P3.13. Find the Jones vector representation for the polarization of the reflected light.

NOTE: The answer may be somewhat different than the result measured in L 6.9. For one thing, we have not considered that a silver mirror inevitably has a thin oxide layer or, more often, a special protective coating applied.

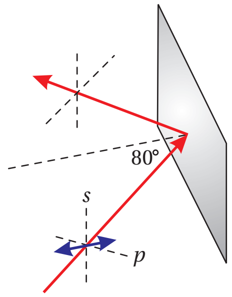


Figure 6.14 Geometry for P6.13.

P6.14 Calculate the angle θ to cut the glass in a Fresnel rhomb such that after the two internal reflections there is a phase difference of $\pi/2$ between the two polarization states. The rhomb then acts as a quarter-wave plate.

HINT: You need to find the phase difference between (3.42) and (3.43). Set the difference equal to $\pi/4$ for each bounce. The equation you get does not have a clean analytic solution, but you can plot it to find a numerical solution.

Answer: There are two angles that work: $\theta \cong 50°$ and $\theta \cong 53°$.

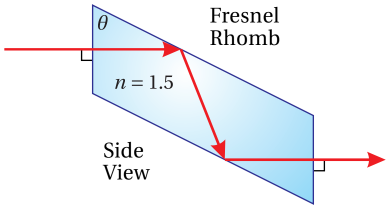


Figure 6.15 Fresnel Rhomb geometry for P6.14.

Exercises for 6.A Ellipsometry

P6.15 Derive (6.49) and (6.51), often used for ellipsometry measurements.

HINT: Using $\sin^2\theta = \frac{1-\cos 2\theta}{2}$ and $\cos^2\theta = \frac{1+\cos 2\theta}{2}$, first show

$$I \propto 1 - \frac{\frac{r_p r_s^* + r_s r_p^*}{|r_s|^2}\tan\alpha}{\frac{|r_p|^2}{|r_s|^2} + \tan^2\alpha}\sin 2\theta + \frac{\frac{|r_p|^2}{|r_s|^2} - \tan^2\alpha}{\frac{|r_p|^2}{|r_s|^2} + \tan^2\alpha}\cos 2\theta$$

Exercises for 6.B Partially Polarized Light

P6.16 Derive the Mueller matrix for a quarter-wave plate.

Answer:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos^2 2\theta & \frac{1}{2}\sin 4\theta & -\sin 2\theta \\ 0 & \frac{1}{2}\sin 4\theta & \sin^2 2\theta & \cos 2\theta \\ 0 & \sin 2\theta & -\cos 2\theta & 0 \end{bmatrix}$$

Chapter 7

Superposition of Quasi-Parallel Plane Waves

In previous chapters, we considered only individual plane-wave fields which have uniform intensity throughout space and time. Some optical fields can be well-approximated by a plane wave, but most have a more complicated structure. It turns out that any field (e.g. pulses or a focused beam), regardless of how complicated, can be described by a superposition of many plane wave fields. In this chapter, we develop the techniques for superimposing plane waves.

We begin our analysis with a discrete sum of plane wave fields and show how to calculate the intensity in this case. We will introduce the concept of *group velocity*, which describes the motion of interference 'fringes' or 'packets' resulting when multiple plane waves are superimposed. Group velocity is distinct from *phase velocity* that we encountered previously. As we saw in chapter 2, the real part of refractive index in certain situations can be less than one, indicating *superluminal* wave crest propagation (i.e. greater than c)! However, it is the group velocity that tracks the speed of interference fringes, which are associated with light intensity.

In section 7.3, we extend our analysis of wave superposition to a continuum of plane waves. The analysis is based on Fourier theory, which is a tool for keeping track of the plane waves that make up a given waveform $\mathbf{E}(\mathbf{r}, t)$. We will learn how to decompose arbitrary waveforms into plane wave components, which we know how to propagate in a material (with a frequency-dependent index). Conversely, we will also learn how to reassemble plane waves into a final pulse at the end of propagation.

Different frequency components of a waveform experience different phase velocities, causing the waveform to undergo distortion as it propagates, a phenomenon called *dispersion*. *Narrowband* packets (i.e. packets comprised of a narrow range of frequencies and hence long duration) tend to maintain their shape (with some spreading) while propagating at the group velocity. On the other hand, *broadband* pulses (i.e. packets comprised of a wide range of frequencies and possibly of short duration) tend to distort severely while propagating in materials.

It turns out that group velocity can also become superluminal when significant absorption and/or amplification of the light pulse is involved. This is no cause for alarm (nor is it cause for an abundance of gee-whiz papers on the subject). Absorption and amplification can cause a pulse to appear to move unexpectedly fast through a reshaping effect. Group velocity, or rather its inverse *group delay*, takes this reshaping into account. For example, energy can be lost from the back of a pulse or perhaps added to an already-present forward portion of a pulse such that the average pulse position appears to advance superluminally. When all energy is accounted for (both the energy in the medium and in the light pulse), however, no information advances faster than the universal speed limit c. Appendix 7.B provides analysis of how a medium exchanges energy with a pulse to produce these eye-catching effects.

7.1 Intensity of Superimposed Plane Waves

We can construct arbitrary waveforms by adding together many plane waves with different propagation directions, amplitudes, phases, frequencies and polarizations. Consider the following discrete sum of plane waves:

$$\mathbf{E}(\mathbf{r}, t) = \sum_j \mathbf{E}_j e^{i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)} \tag{7.1}$$

The corresponding magnetic field according to (2.56) is

$$\mathbf{B}(\mathbf{r}, t) = \sum_j \mathbf{B}_j e^{i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)} = \sum_j \frac{\mathbf{k}_j \times \mathbf{E}_j}{\omega_j} e^{i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)} \tag{7.2}$$

As usual, the (time- and space-independent) individual field components $\mathbf{E}_j$ contain both amplitude and phase information for each plane wave.

The Poynting vector (2.52) associated with the fields (7.1) and (7.2) is

$$\begin{aligned}\mathbf{S}(\mathbf{r}, t) &= \mathrm{Re}\{\mathbf{E}(\mathbf{r}, t)\} \times \frac{\mathrm{Re}\{\mathbf{B}(\mathbf{r}, t)\}}{\mu_0} \\ &= \sum_{j,m} \frac{1}{\omega_m \mu_0} \mathrm{Re}\left\{\mathbf{E}_j e^{i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)}\right\} \times \mathrm{Re}\left\{\mathbf{k}_m \times \mathbf{E}_m e^{i(\mathbf{k}_m\cdot\mathbf{r}-\omega_m t)}\right\}\end{aligned} \tag{7.3}$$

(Recall the conspiracy that only the real parts of the fields are relevant – crucial before multiplying.) The above expression is cumbersome because of the many cross terms that arise when the two summations are multiplied. We need some simplifying assumptions before we can make any real progress on this expression. For example, we can time-average the Poynting vector to remove fluctuations that vary on the scale of optical frequencies. Additionally, it is common to encounter the situation where all plane-wave components travel roughly parallel to each other, which will be a big help in simplifying (7.3). Let us further assume that the $\mathbf{k}_m$ vectors are real.[1]

Intensity for Quasi Parallel-traveling Light

We apply the BAC-CAB rule (P0.3) to (7.3) and obtain

$$\begin{aligned}\mathbf{S}(\mathbf{r},t) = \sum_{j,m} \frac{1}{\omega_m \mu_0} \Big[& \mathbf{k}_m \left(\mathrm{Re}\left\{ \mathbf{E}_j e^{i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)} \right\} \cdot \mathrm{Re}\left\{ \mathbf{E}_m e^{i(\mathbf{k}_m\cdot\mathbf{r}-\omega_m t)} \right\} \right) \\ & - \mathrm{Re}\left\{ \mathbf{E}_m e^{i(\mathbf{k}_m\cdot\mathbf{r}-\omega_m t)} \right\} \left(\mathrm{Re}\left\{ \mathbf{E}_j e^{i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)} \right\} \cdot \mathbf{k}_m \right) \Big]\end{aligned} \tag{7.4}$$

The last term in (7.4) can be dismissed if all $\mathbf{k}$-vectors are approximately parallel to each other, in which case all of the $\mathbf{k}_m$ are essentially perpendicular to each of the $\mathbf{E}_j$. We will make this rather stringent assumption and kill the last line in (7.4). The magnitude of the Poynting vector then becomes (with the help of (0.30))

$$\begin{aligned} S(\mathbf{r},t) &= \sum_{j,m} \frac{k_m}{\omega_m \mu_0} \left\{ \frac{\mathbf{E}_j e^{i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)} + \mathbf{E}_j^* e^{-i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)}}{2} \right\} \\ &\qquad\qquad \cdot \left\{ \frac{\mathbf{E}_m e^{i(\mathbf{k}_m\cdot\mathbf{r}-\omega_m t)} + \mathbf{E}_m^* e^{-i(\mathbf{k}_m\cdot\mathbf{r}-\omega_m t)}}{2} \right\} \qquad \text{(parallel } \mathbf{k}\text{-vectors)} \\ &= \sum_{j,m} \frac{k_m}{4\omega_m \mu_0} \Big\{ \mathbf{E}_j \cdot \mathbf{E}_m e^{i[(\mathbf{k}_j+\mathbf{k}_m)\cdot\mathbf{r}-(\omega_j+\omega_m)t]} + \mathbf{E}_j^* \cdot \mathbf{E}_m^* e^{-i[(\mathbf{k}_j+\mathbf{k}_m)\cdot\mathbf{r}-(\omega_j+\omega_m)t]} \\ &\qquad + \mathbf{E}_j \cdot \mathbf{E}_m^* e^{i[(\mathbf{k}_j-\mathbf{k}_m)\cdot\mathbf{r}-(\omega_j-\omega_m)t]} + \mathbf{E}_j^* \cdot \mathbf{E}_m e^{-i[(\mathbf{k}_j-\mathbf{k}_m)\cdot\mathbf{r}-(\omega_j-\omega_m)t]} \Big\} \end{aligned} \tag{7.5}$$

The terms involving $(\omega_j + \omega_m)t$ oscillate rapidly and time-average to zero. By comparison, the terms involving $(\omega_j - \omega_m)t$ oscillate slowly (especially when the ω_j are all in the neighborhood of the ω_m) or not at all when $j = m$. We retain the slower fluctuations and discard the rapid oscillations. For purposes of computing the intensity we can approximate the index as approximately constant, and write $k_m/(\omega_m\mu_0) \approx n\epsilon_0 c$. With these simplifications, (7.5) becomes

$$\begin{aligned} \langle S(\mathbf{r},t)\rangle_{\text{osc}} &= \frac{n\epsilon_0 c}{2} \sum_{j,m} \frac{\mathbf{E}_j \cdot \mathbf{E}_m^* e^{i[(\mathbf{k}_j-\mathbf{k}_m)\cdot\mathbf{r}-(\omega_j-\omega_m)t]} + \mathbf{E}_j^* \cdot \mathbf{E}_m e^{-i[(\mathbf{k}_j-\mathbf{k}_m)\cdot\mathbf{r}-(\omega_n-\omega_m)t]}}{2} \\ &= \frac{n\epsilon_0 c}{2} \mathrm{Re}\left\{ \sum_j \mathbf{E}_j e^{i(\mathbf{k}_j\cdot\mathbf{r}-\omega_j t)} \cdot \sum_m \mathbf{E}_m^* e^{-i(\mathbf{k}_m\cdot\mathbf{r}-\omega_m t)} \right\} \\ &= \frac{n\epsilon_0 c}{2} \mathrm{Re}\left\{ \mathbf{E}(\mathbf{r},t) \cdot \mathbf{E}^*(\mathbf{r},t) \right\}. \end{aligned}$$

$$\text{(parallel } \mathbf{k}\text{-vectors)} \tag{7.6}$$

The final expression in (7.6) is already manifestly real so there is no need to apply the operation Re{ }. The time-averaged intensity for light composed of parallel wave vectors is then well-approximated by

---

[1] If the wave vectors are complex, the result is essentially the same, but, as in (2.62), the field amplitudes $\mathbf{E}_j$ correspond to local amplitudes (adjusted for absorption or amplification during prior propagation).

(valid for parallel or antiparallel **k**-vectors and approximately constant n)

$$I(\mathbf{r}, t) = \frac{n\epsilon_0 c}{2}\mathbf{E}(\mathbf{r}, t)\cdot\mathbf{E}^*(\mathbf{r}, t) \tag{7.7}$$

In a surprising turn of events, it is important that $\mathbf{E}(\mathbf{r}, t)$ in (7.7) be written as the entire complex expression for the electric field rather than just the real part. Then (7.7) automatically time-averages over rapid oscillations in such a way that $I(\mathbf{r}, t)$ retains a *slowly varying* time dependence. This expression is reminiscent of (2.62), but it should be kept in mind that we previously considered only a single plane wave (perhaps with two distinct polarization components).

If some of the **k**-vectors point in an anti-parallel direction, we can still use (7.7). This brings up a distinction between irradiance S and intensity I. For example, $\langle S\rangle$ is zero for standing waves because there is no net flow of energy, whereas (7.7) still gives a result. Intensity specifies whether atoms locally experience an oscillating electric field without regard for whether there is a net flow of energy carried by a light field.[2]

We can relax the restriction of parallel $\mathbf{k}_j$'s slightly and apply (7.7) also to plane waves with *nearly* parallel $\mathbf{k}_j$'s. Such a situation occurs, for example, in a Young's two-slit diffraction experiment (studied in chapter 8).

7.2 Group vs. Phase Velocity: Sum of Two Plane Waves

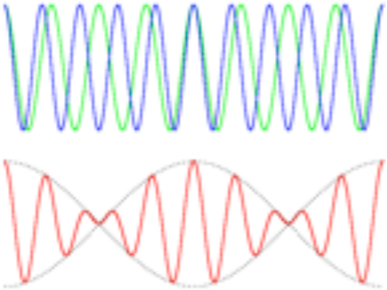


Figure 7.1 Animation showing superposition of two plane waves (electric fields) with different frequencies and traveling at different speeds.

To begin our study of interference, consider just two plane waves with equal amplitudes given by

$$\mathbf{E}_1 = \mathbf{E}_0 e^{i(\mathbf{k}_1\cdot\mathbf{r}-\omega_1 t)} \quad \text{and} \quad \mathbf{E}_2 = \mathbf{E}_0 e^{i(\mathbf{k}_2\cdot\mathbf{r}-\omega_2 t)} \tag{7.8}$$

As we previously studied (see P1.9), the velocities of the wave crests for these two waves are

$$v_{p1} = \omega_1/k_1 \quad \text{and} \quad v_{p2} = \omega_2/k_2 \tag{7.9}$$

These are known as the *phase velocities* of the individual plane waves.

Next consider a composite wave created from the superposition of the above two plane waves:

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 e^{i(\mathbf{k}_1\cdot\mathbf{r}-\omega_1 t)} + \mathbf{E}_0 e^{i(\mathbf{k}_2\cdot\mathbf{r}-\omega_2 t)} \tag{7.10}$$

The two plane waves interfere, producing regions of higher and lower intensity that move in time. Remarkably, these intensity peaks can propagate at speeds quite different from either of the phase velocities in (7.9). The intensity (7.7) for the field (7.10) is computed as follows:

---

[2] At extreme intensities, when the influence of the magnetic field becomes comparable to that of the electric field, the distinction between propagating and standing fields becomes important to the behavior of charged particles in that field.

$$\begin{aligned} I(\mathbf{r},t) &= \frac{n\epsilon_0 c}{2}\mathbf{E}_0\cdot\mathbf{E}_0^*\left[e^{i(\mathbf{k}_1\cdot\mathbf{r}-\omega_1 t)}+e^{i(\mathbf{k}_2\cdot\mathbf{r}-\omega_2 t)}\right]\left[e^{-i(\mathbf{k}_1\cdot\mathbf{r}-\omega_1 t)}+e^{-i(\mathbf{k}_2\cdot\mathbf{r}-\omega_2 t)}\right] \\ &= \frac{n\epsilon_0 c}{2}\mathbf{E}_0\cdot\mathbf{E}_0^*\left[2+e^{i[(\mathbf{k}_2-\mathbf{k}_1)\cdot\mathbf{r}-(\omega_2-\omega_1)t]}+e^{-i[(\mathbf{k}_2-\mathbf{k}_1)\cdot\mathbf{r}-(\omega_2-\omega_1)t]}\right] \\ &= n\epsilon_0 c\mathbf{E}_0\cdot\mathbf{E}_0^*\left[1+\cos\left[(\mathbf{k}_2-\mathbf{k}_1)\cdot\mathbf{r}-(\omega_2-\omega_1)t\right]\right] \\ &= n\epsilon_0 c\mathbf{E}_0\cdot\mathbf{E}_0^*\left[1+\cos(\Delta\mathbf{k}\cdot\mathbf{r}-\Delta\omega t)\right] \end{aligned} \tag{7.11}$$

where

$$\begin{aligned} \Delta\mathbf{k} &\equiv \mathbf{k}_2-\mathbf{k}_1 \\ \Delta\omega &\equiv \omega_2-\omega_1 \end{aligned} \tag{7.12}$$

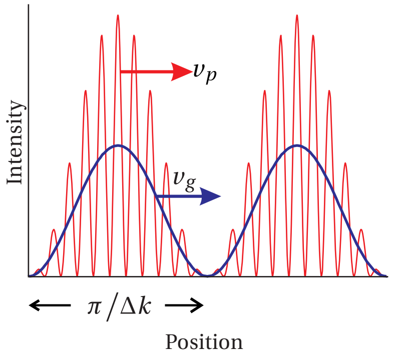


Figure 7.2 Intensity of two interfering plane waves. The solid line shows intensity averaged over rapid oscillations.

The darker line in Fig. 7.2 shows the intensity computed with (7.11). Keep in mind that this intensity is averaged over rapid oscillations. For comparison, the lighter line shows the Poynting flux with the rapid oscillations retained, according to (7.5). It is left as an exercise (see P7.3) to show that the rapid-oscillation peaks in Fig. 7.2 move with a phase velocity derived from the average **k** and average ω of the two plane waves:

$$v_p \equiv \frac{\bar{\omega}}{\bar{k}} \tag{7.13}$$

An examination of the cosine argument in (7.11) reveals that the time-averaged curve in Fig. 7.2 (dark) travels with speed

$$v_g \equiv \frac{\Delta\omega}{\Delta k} \cong \left.\frac{d\omega}{dk}\right|_{\bar{\omega}} \tag{7.14}$$

This is known as the *group velocity*. Essentially, v_g may be thought of as the velocity for the envelope that encloses the rapid oscillations. As noted, the group velocity is often written as a derivative rather than a ratio of finite differences; the derivative will be more natural when dealing with a continuum of plane waves rather than a pair of planes.

In general, v_g and v_p are not the same. This means that as the waveform propagates, the rapid oscillations move within the larger modulation pattern, for example, continually disappearing at the front and reappearing at the back of each modulation. The group velocity is identified with the propagation of overall waveforms. The presence of intensity in a waveform is clearly tied more to v_g than to v_p.

John William Strutt (3rd Baron Rayleigh) (1842–1919, British) was born in Langford Grove, Essex, England and was frequently ill in his youth. He entered the University of Cambridge in 1861 and graduated four years later as senior wrangler in mathematics. He married in 1871 and became the father of three sons. In 1873, Strutt inherited the Barony of Rayleigh (and the title Lord Rayleigh) from his father who died that year. In 1879 Strutt succeeded James Clerk Maxwell as the Cavendish Professor of Physics at Cambridge. Rayleigh studied a wide variety of subjects. He is credited with the discovery of argon. He studied how atoms scatter light (Rayleigh scattering) and explained why the sky is blue. He extensively developed the notion of group velocity and used it to understand the propagation of sound. He won the Nobel prize in physics in 1904 for investigations of gas densities and for discovering argon. (Wikipedia)

Example 7.1

Determine the phase velocity and group velocity for the superposition of two plane waves in a plasma (see P2.7).

Solution: The index of refraction is given by

$$n_{\text{plasma}}(\omega) = \sqrt{1-\omega_{\text{p}}^2/\omega^2} < 1 \qquad \text{(assuming } \omega > \omega_{\text{p}}\text{)} \tag{7.15}$$

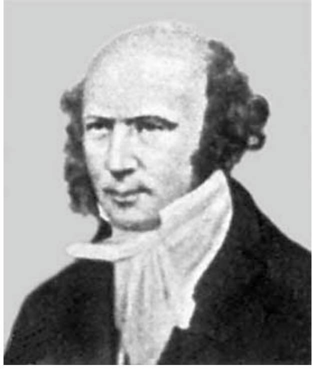
The phase velocity (7.13) is computed as

$$v_p = \frac{\omega_1 + \omega_2}{n_{\text{plasma}}(\omega_1)\omega_1/c + n_{\text{plasma}}(\omega_2)\omega_2/c} \cong \frac{c}{n_{\text{plasma}}(\omega)} \tag{7.16}$$

For convenience, we have taken ω_1 and ω_2 to lie very close to each other. Since $n_{\text{plasma}} < 1$, (7.16) shows that the phase velocity exceeds c. However, the group velocity is

$$v_g = \frac{\Delta\omega}{\Delta k} \cong \frac{d\omega}{dk} = \left[\frac{dk}{d\omega}\right]^{-1} = \left[\frac{d}{d\omega}\frac{\omega n_{\text{plasma}}(\omega)}{c}\right]^{-1} = n_{\text{plasma}}(\omega)\, c \tag{7.17}$$

which is clearly less than c. The derivation of the final expression in (7.17) from the previous one is left as an exercise.



Sir William Rowen Hamilton (1805–1865, Irish) was born in Dublin, Ireland, the fourth of nine children. At a very early age, he showed a remarkable ability to learn languages while living with his uncle who was a linguist. He became proficient in nearly a dozen languages and in later life enjoyed reading in various languages as a means of relaxation. At age eight, Hamilton entered a mental arithmetic contest against a nine-year-old prodigy from America. Hamilton lost and as a result determined to spend much more time on mathematics instead of languages. Hamilton went on to make enormous contributions to mathematical physics. His reformulation of classical dynamics proved to be the ideal framework for later developments in electrodynamics, quantum mechanics, and quantum field theory. Ironically, Hamilton was originally employed as an observational astronomer at Dunsink Obervatory, a post for which he was not particularly well suited. The University of Dublin didn't mind, however, owing to the outstanding quality of his theoretical pursuits. Hamilton is credited with first articulating the concept of group velocity, although only abstracts of his lectures on the subject have been preserved: *Researches respecting vibration, connected with the theory of light*, Proc. Roy. Irish Acad. 1, 267, 341 (1839).(Wikipedia)

Example 7.1 illustrates that in an environment where the index of refraction is real (i.e. no net exchange of energy with the medium), the group velocity does not exceed c, even when the phase velocity does. The 'fast-moving' phase velocity v_p results merely from an interplay between the field and the plasma. In a similar sense, the intersection of an ocean wave with the shoreline can also exceed c, if different points on the wave front happen to strike the shore nearly simultaneously. The point of intersection between the wave and the shoreline does not constitute an actual object under motion. Similarly, wave crests of a wave, at least when interacting with a medium, do not necessarily constitute actual objects in motion. That is, v_p is not the relevant speed at which events upstream influence events downstream in a medium.

7.3 Frequency Spectrum of Light

Individual plane waves have infinite length and infinite duration. They do not exist in isolation except in our imagination. Moreover, a waveform constructed from a discrete sum (as in the previous two sections) must eventually repeat over and over (i.e. it is periodic). To create a waveform that does not repeat (e.g. a single laser pulse or, technically speaking, any waveform that exists in the physical world since no light source repeats forever) we must replace the discrete sum (7.1) with an integral that combines a continuum of plane waves. Such a waveform at a point $\mathbf{r}$ can be expressed as

$$\mathbf{E}(\mathbf{r}, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbf{E}(\mathbf{r}, \omega)\, e^{-i\omega t} d\omega \tag{7.18}$$

The function $\mathbf{E}(\mathbf{r}, \omega)$, called the *spectrum*, has units of *field per frequency*. Essentially, it gives the amplitude and phase of each plane wave that makes up the overall waveform. It includes any spatially dependent factors such as $\exp\{i\mathbf{k}(\omega)\cdot\mathbf{r}\}$. We distinguish the spectrum $\mathbf{E}(\mathbf{r}, \omega)$ from the wholly separate function $\mathbf{E}(\mathbf{r}, t)$ by its argument (i.e. ω instead of t). (Sorry for using $\mathbf{E}$ for both functions, but this is standard notation.) The operation (7.18) is called an *inverse Fourier transform* as outlined in section 0.4; it would be a good idea to review section 0.4 thoroughly. Now. Why haven't you turned to section 0.4 yet? The factor $1/\sqrt{2\pi}$ is introduced to match our Fourier-transform convention. Notice that (7.18) merely sums together a range of plane waves in much the same way that the discrete summation (7.1) does.

Given a waveform $\mathbf{E}(\mathbf{r},t)$, one might wonder what plane waves should be added together in order to construct it. Equation (7.18) can be inverted, which remarkably has a very similar form:

$$\mathbf{E}(\mathbf{r},\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathbf{E}(\mathbf{r},t)\, e^{i\omega t} dt \tag{7.19}$$

This operation is called the *Fourier transform.* It is used to generate the spectrum $\mathbf{E}(\mathbf{r},\omega)$ from the field $\mathbf{E}(\mathbf{r},t)$ in much the same way that (7.18) is used to generate the field $\mathbf{E}(\mathbf{r},t)$ from the spectrum $\mathbf{E}(\mathbf{r},\omega)$.

Although only the real part of $\mathbf{E}(\mathbf{r},t)$ is physically relevant, we can continue our habit of working with the complex field and taking the real part of $\mathbf{E}(\mathbf{r},t)$ at our leisure.[3] In fact, we will find it advantageous to work with the complex field instead of only the real part. We will not run into trouble as long as we remember never to discard the imaginary part of $\mathbf{E}(\mathbf{r},\omega)$, only the imaginary part of $\mathbf{E}(\mathbf{r},t)$.

The intensity formula (7.7) remains useful for continuous superpositions of plane waves (i.e. a field defined by the inverse Fourier transform (7.18)):

$$I(\mathbf{r},t) \equiv \frac{n\epsilon_0 c}{2}\mathbf{E}(\mathbf{r},t)\cdot\mathbf{E}^*(\mathbf{r},t) \tag{7.20}$$

Remember, this formula specifically requires the fields to be in complex format, and it takes care of the time-average over rapid oscillations automatically.[4] Moreover, the above expression for $I(\mathbf{r},t)$ assumes that all relevant $\mathbf{k}$-vectors are essentially parallel.

Similarly, we will define the *power spectrum* produced from $\mathbf{E}(\mathbf{r},\omega)$, which we write as

$$I(\mathbf{r},\omega) \equiv \frac{n\epsilon_0 c}{2}\mathbf{E}(\mathbf{r},\omega)\cdot\mathbf{E}^*(\mathbf{r},\omega) \tag{7.21}$$

The power spectrum $I(\mathbf{r},\omega)$ is what one observes when the waveform is sent into a spectral analyzer or spectrometer. We must apologize again for the potentially confusing notation (in wide usage): $I(\mathbf{r},\omega)$ is not the Fourier transform of $I(\mathbf{r},t)$! They are defined exclusively through (7.20) and (7.21).

---

[3] Since Fourier transforms are linear, one can take the Fourier transform of the real and imaginary parts of a field separately. Appropriate modifications to $\mathbf{E}(\mathbf{r},\omega)$ in the frequency domain will not cause the two parts to become mingled. Upon taking the inverse Fourier transform to obtain $\mathbf{E}(\mathbf{r},t)$ again, the original real part remains purely real, and the original imaginary part remains purely imaginary.

[4] To use this expression there needs to be a sufficient number of oscillations within the waveform to make the rapid time average meaningful.

Parseval's theorem (see Example 0.7) imposes an interesting connection between the time-integral of the intensity and the frequency-integral of the power spectrum:

$$\int_{-\infty}^{\infty} I(\mathbf{r},t)\,dt = \int_{-\infty}^{\infty} I(\mathbf{r},\omega)\,d\omega \tag{7.22}$$

With the above formalities out of the way, we will illustrate the use of Fourier transforms through some examples.

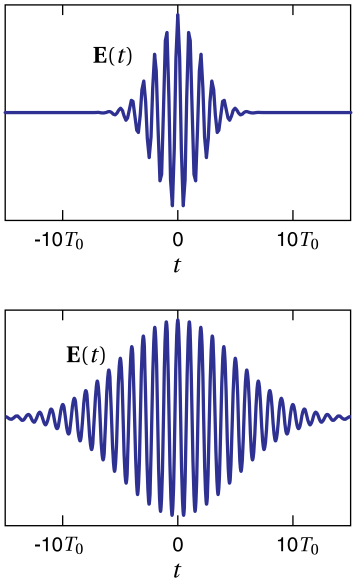


Figure 7.3 Real part of electric field (7.23) with $T = 2T_0$ and $T = 5T_0$, where $T_0 = 2\pi/\omega_0$ is the period of the carrier frequency.

Example 7.2

Find $\mathbf{E}(\mathbf{r},\omega)$ associated with the field

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0(\mathbf{r})\,e^{-t^2/2T^2}\,e^{-i\omega_0 t} \tag{7.23}$$

The real part of this field is shown in Fig. 7.3 for two different durations T. The intensity profile computed by (7.20) is shown in Fig. 7.4 .

Solution: The argument $\mathbf{r}$ is unimportant to our calculation. It merely specifies that we are considering the field at the point $\mathbf{r}$. We compute the Fourier transform as follows:

$$\begin{aligned}\mathbf{E}(\mathbf{r},\omega) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathbf{E}_0(\mathbf{r})\,e^{-t^2/2T^2}\,e^{-i\omega_0 t}e^{i\omega t}dt\\ &= \frac{\mathbf{E}_0(\mathbf{r})}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-t^2/2T^2+i(\omega-\omega_0)t}dt\end{aligned} \tag{7.24}$$

This integral can be performed with the help of (0.55), and we obtain

$$\mathbf{E}(\mathbf{r},\omega) = T\mathbf{E}_0(\mathbf{r})\,e^{-\frac{T^2(\omega-\omega_0)^2}{2}} \tag{7.25}$$

Notice that $\mathbf{E}(\mathbf{r},\omega)$ has units of field multiplied by time, or in other words, field per frequency.

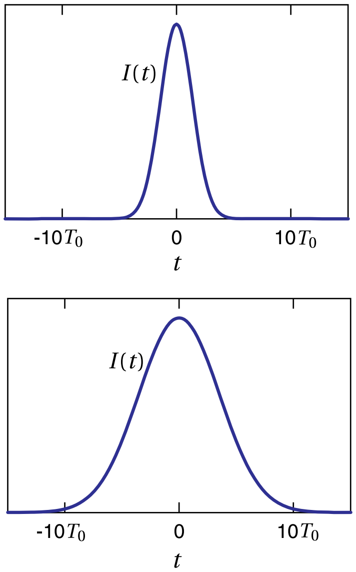


Figure 7.4 The intensity (7.20) of the fields in Fig. 7.3.

In general, $\mathbf{E}(\mathbf{r},\omega)$ is a complex function. $\mathbf{E}(\mathbf{r},\omega)$ keeps track of the amplitude and phase of each plane wave needed to compose the waveform $\mathbf{E}(\mathbf{r},t)$. More often than not, $\mathbf{E}(\mathbf{r},\omega)$ exhibits a complicated complex phase structure, depending on the time-shape of $\mathbf{E}(\mathbf{r},t)$.

The spectrum of the field in Example 7.2 is shown in Fig. 7.5. The complex phase turns out to be boringly uniform for this example; if $\mathbf{E}_0$ is real, the imaginary part of the spectrum turns out to be zero for all frequencies. The corresponding power spectrum (7.21) is plotted in Fig. 7.6. As expected, the waveform includes frequencies in the neighborhood of ω_0.

A range of frequencies are needed to construct a waveform that turns on and off. The shorter the duration of the waveform, the wider the frequency spectrum that is necessary. Note that the temporal width of the waveform (7.23) is dictated by T while the spectral width of (7.25) is given by $\Omega \equiv 1/T$. This gives an uncertainty product $T\Omega = 1$. This dictates the minimum spectral width necessary to produce a pulse of a given duration.

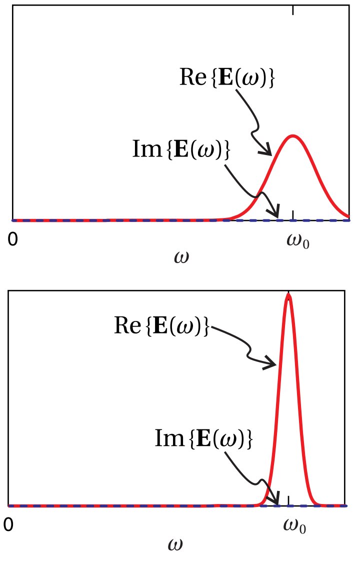


Figure 7.5 Spectral components (7.25) of the fields in Fig. 7.3 with $T = 4\pi/\omega_0$ and $T = 10\pi/\omega_0$, where $2\pi/\omega_0$ is the period of the carrier frequency.

Example 7.3

Check Parseval's theorem for the field and spectrum in Example 7.2.

Solution: The time integration in (7.22) yields

$$\begin{aligned}\int_{-\infty}^{\infty} I(\mathbf{r}, t)\,dt &= \frac{n\epsilon_0 c}{2}\mathbf{E}_0(\mathbf{r})\cdot\mathbf{E}_0^*(\mathbf{r})\int_{-\infty}^{\infty} e^{-t^2/T^2}\,dt\\ &= \frac{n\epsilon_0 c}{2}\mathbf{E}_0(\mathbf{r})\cdot\mathbf{E}_0^*(\mathbf{r})\,T\sqrt{\pi}\end{aligned}$$

where we have used (0.55) to perform the integration. This result has units of energy per area, called *fluence*. It is the total energy per area absorbed by a detector over the entire pulse. The frequency integration in (7.22) yields

$$\begin{aligned}\int_{-\infty}^{\infty} I(\mathbf{r}, \omega)\,d\omega &= \frac{n\epsilon_0 c}{2}\mathbf{E}_0(\mathbf{r})\cdot\mathbf{E}_0^*(\mathbf{r})\,T^2\int_{-\infty}^{\infty} e^{-T^2(\omega-\omega_0)^2}\,d\omega\\ &= \frac{n\epsilon_0 c}{2}\mathbf{E}_0(\mathbf{r})\cdot\mathbf{E}_0^*(\mathbf{r})\,T^2\frac{\sqrt{\pi}}{T}\end{aligned}$$

which is the same answer.

As mentioned previously, the inverse Fourier transform is interpreted as summing together many plane waves to create a waveform.

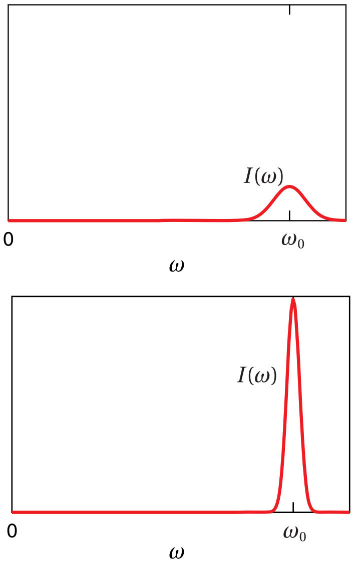


Figure 7.6 Power spectrum based on (7.21) for the spectral components shown in Fig. 7.5.

Example 7.4

Take the inverse Fourier transform of (7.25) to recover the original waveform (7.23).

Solution: The inverse Fourier transform (7.18) is

$$\begin{aligned}\mathbf{E}(\mathbf{r}, t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathbf{E}(\mathbf{r}, \omega)\,e^{-i\omega t}\,d\omega\\ &= \frac{T\mathbf{E}_0(\mathbf{r})}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{T^2(\omega-\omega_0)^2}{2}}e^{-i\omega t}\,d\omega\\ &= \frac{T\mathbf{E}_0(\mathbf{r})}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{T^2\omega^2}{2}+(T^2\omega_0 - it)\omega-\frac{T^2\omega_0{}^2}{2}}\,d\omega\end{aligned} \tag{7.26}$$

This integral can be performed with the help of (0.55), which gives

$$\begin{aligned}\mathbf{E}(\mathbf{r}, t) &= \frac{T\mathbf{E}_0(\mathbf{r})}{\sqrt{2\pi}}\sqrt{\frac{\pi}{T^2/2}}\,e^{\frac{(T^2\omega_0 - it)^2}{4(T^2/2)}-\frac{T^2\omega_0^2}{2}}\\ &= \mathbf{E}_0(\mathbf{r})\,e^{-t^2/2T^2}\,e^{-i\omega_0 t}\end{aligned}$$

Since only the real part of the time profile $\mathbf{E}(\mathbf{r}, t)$ is physically relevant, you might be curious about how the Fourier transform of the real part of the field compares with that of the complex version of the field that we have been using. Indeed, there are situations where it is more appropriate to use the real version of the field rather than its complex form. For example, if a waveform includes multiple propagation directions or if a waveform contains only a few cycles, then the motivation/interpretation behind (7.20) and the convenience of the complex format begin to wane.

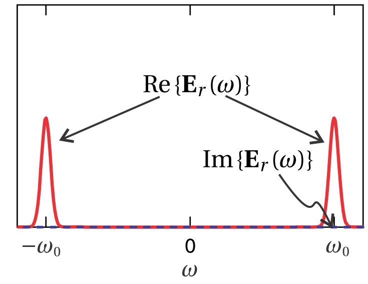


Figure 7.7 Spectrum based on (7.28) with $T = 10\pi/\omega_0$. Compare with the lower curve in Fig. 7.5

Example 7.5

Take the Fourier transform of just the real part of waveform (7.23).

Solution: The real part of (7.23) is

$$\begin{aligned}\mathbf{E}_r(\mathbf{r}, t) &= \frac{\mathbf{E}(\mathbf{r}, t) + \mathbf{E}^*(\mathbf{r}, t)}{2} \\ &= e^{-t^2/2T^2}\,\frac{\mathbf{E}_0(\mathbf{r})\, e^{-i\omega_0 t} + \mathbf{E}_0^*(\mathbf{r})\, e^{i\omega_0 t}}{2}\end{aligned} \tag{7.27}$$

If $\mathbf{E}_0(\mathbf{r})$ is real, then this field can be written as $\mathbf{E}_0(\mathbf{r})\, e^{-t^2/2T^2}\cos(\omega_0 t)$. The Fourier transform (7.19) yields (see P0.24)

$$\mathbf{E}_r(\mathbf{r}, \omega) = T\,\frac{\mathbf{E}_0(\mathbf{r})\, e^{-\frac{T^2(\omega+\omega_0)^2}{2}} + \mathbf{E}_0^*(\mathbf{r})\, e^{-\frac{T^2(\omega-\omega_0)^2}{2}}}{2} \tag{7.28}$$

The spectrum is shown in Fig. 7.7.

From the above example, you might notice that the transform of the real part of a field tends to be more cumbersome than the transform of the entire complex field. For the real field, both positive and negative frequency components contribute to the overall spectrum.[5] Moreover, the Fourier transform of a real function $\mathbf{E}_r(\mathbf{r}, t)$ obeys the symmetry relation

$$\mathbf{E}_r(\mathbf{r}, -\omega) = \mathbf{E}_r^*(\mathbf{r}, \omega) \qquad \text{(if } \mathbf{E}_r(\mathbf{r}, t) \text{ is real)} \tag{7.29}$$

whereas the Fourier transform of the complex field depicted in Fig. 7.5 does not.

7.4 Wave Packet Propagation and Group Delay

Once we have the spectrum for a waveform (obtained by Fourier transform), we can apply effects to the individual spectral components. In particular, we can find how an overall waveform propagates in a uniform medium by taking advantage of our knowledge of how individual plane waves propagate (as studied in chapter 2). At any point in the medium, we can perform an inverse Fourier transform, which recombines spectral components (i.e. plane waves) to reveal how the overall waveform looks as a function of time. Thus, we will be able to predict the temporal profile of a waveform at any location given knowledge of that waveform at another location.[6]

---

[5] Essentially, the spectrum of the complex representation of the field can be understood to be twice the spectrum of the real representation, but plotted only for the positive frequencies.

Let $\mathbf{E}(\mathbf{r}_0, t)$ be the temporal profile of a pulse at some point $\mathbf{r}_0$ in a medium. The spectrum of this pulse $\mathbf{E}(\mathbf{r}_0, \omega)$ (found using (7.19)) gives the amplitudes and phases of the individual plane wave components at the point $\mathbf{r}_0$. A phase shift associated with a displacement $\Delta\mathbf{r}$ modifies the spectral components according to (see (2.20))

$$\mathbf{E}(\mathbf{r}_0 + \Delta\mathbf{r}, \omega) = \mathbf{E}(\mathbf{r}_0, \omega)\, e^{i\mathbf{k}(\omega)\cdot\Delta\mathbf{r}} \tag{7.30}$$

The $\mathbf{k}$-vector contains the frequency-dependent information about the material via $k = n(\omega)\omega/c$.[7] We take the inverse Fourier transform of $\mathbf{E}(\mathbf{r}_0 + \Delta\mathbf{r}, \omega)$ at the new position to determine the waveform $\mathbf{E}(\mathbf{r}_0 + \Delta\mathbf{r}, t)$:

$$\begin{aligned}
\mathbf{E}(\mathbf{r}_0 + \Delta\mathbf{r}, t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbf{E}(\mathbf{r}_0 + \Delta\mathbf{r}, \omega) e^{-i\omega t}\, d\omega \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbf{E}(\mathbf{r}_0, \omega) e^{i(\mathbf{k}(\omega)\cdot\Delta\mathbf{r} - \omega t)}\, d\omega
\end{aligned} \tag{7.31}$$

Example 7.6 If a waveform at $\mathbf{r}_0 = 0$ has the form $\mathbf{E}(0, t) = \mathbf{E}_0 e^{-t^2/2T^2} e^{-i\omega_0 t}$, compute the waveform at $\mathbf{r} = z\hat{\mathbf{z}}$ if propagation occurs in vacuum in the z-direction.

Solution: Of course, after traversing $\Delta\mathbf{r} = z\hat{\mathbf{z}}$ in vacuum, the waveform will look the same, only arriving a time z/c later. We'll demonstrate that the tools described above yield this expected result. The Fourier transform of the Gaussian pulse is given in (7.25):

$$\mathbf{E}(0, \omega) = T\mathbf{E}_0 e^{-\frac{T^2(\omega-\omega_0)^2}{2}}$$

To find the field downstream we invoke (7.30), assuming $\mathbf{k}(\omega) = k_{\text{vac}}(\omega)\hat{\mathbf{z}} = \frac{\omega}{c}\hat{\mathbf{z}}$, which gives the appropriate phase shift for each plane wave component:

$$\mathbf{E}(z, \omega) = \mathbf{E}(0, \omega)\, e^{i\mathbf{k}(\omega)\cdot\Delta\mathbf{r}} = T\mathbf{E}_0 e^{-\frac{T^2(\omega-\omega_0)^2}{2}} e^{i\frac{\omega}{c}z}$$

We compute the final waveform using (7.31) and obtain

$$\mathbf{E}(z, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbf{E}_0 T e^{-\frac{T^2(\omega-\omega_0)^2}{2}} e^{i\frac{\omega}{c}z} e^{-i\omega t}\, d\omega = \mathbf{E}_0 e^{-\frac{(t-z/c)^2}{2T^2}} e^{-i\omega_0(t-z/c)} \tag{7.32}$$

which is the original pulse delayed by z/c.

---

[6] See J. D. Jackson, *Classical Electrodynamics*, 3rd ed., Sect. 7.8 (New York: John Wiley, 1999).

[7] A complex wave vector $\mathbf{k}$ may also be used if absorption or amplification is present.

A waveform propagating in a material such as glass can undergo significant temporal *dispersion*, as different frequency components experience different indices of refraction. Each frequency component propagates at its own phase velocity. The speed of the pulse, however, can be quite different; the pulse as a whole propagates approximately with the group velocity, as will be shown below.

The exponent in (7.30) is called the *phase delay* for the pulse propagation. It is often expanded in a Taylor series about the pulse *carrier frequency* ω_0:

$$\mathbf{k}\cdot\Delta\mathbf{r} \cong \left[\mathbf{k}|_{\omega_0} + \left.\frac{\partial\mathbf{k}}{\partial\omega}\right|_{\omega_0}(\omega-\omega_0) + \frac{1}{2}\left.\frac{\partial^2\mathbf{k}}{\partial\omega^2}\right|_{\omega_0}(\omega-\omega_0)^2 + \cdots\right]\cdot\Delta\mathbf{r} \tag{7.33}$$

The $\mathbf{k}$-vector has a sometimes-complicated frequency dependence through the functional form of $n(\omega)$. If we retain only the first two terms in this expansion then (7.31) becomes

$$\begin{aligned}\mathbf{E}(\mathbf{r}_0+\Delta\mathbf{r},t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathbf{E}(\mathbf{r}_0,\omega)e^{i\left(\left[\mathbf{k}(\omega_0)+\left.\frac{\partial\mathbf{k}}{\partial\omega}\right|_{\omega_0}(\omega-\omega_0)\right]\cdot\Delta\mathbf{r}-\omega t\right)}\,\mathrm{d}\omega \\ &= e^{i\left[\mathbf{k}(\omega_0)-\omega_0\left.\frac{\partial\mathbf{k}}{\partial\omega}\right|_{\omega_0}\right]\cdot\Delta\mathbf{r}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathbf{E}(\mathbf{r}_0,\omega)\,e^{-i\omega\left(t-\left.\frac{\partial\mathbf{k}}{\partial\omega}\right|_{\omega_0}\cdot\Delta\mathbf{r}\right)}\,\mathrm{d}\omega \\ &= e^{i\left[\mathbf{k}(\omega_0)\cdot\Delta\mathbf{r}-\omega_0 t'\right]}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathbf{E}(\mathbf{r}_0,\omega)\,e^{-i\omega(t-t')}\,d\omega\end{aligned} \tag{7.34}$$

where in the last line we have introduced the definition

$$t' \equiv \left.\frac{\partial\mathbf{k}}{\partial\omega}\right|_{\omega_0}\cdot\Delta\mathbf{r} \tag{7.35}$$

The integral in (7.34) is recognized as the Fourier transform of the original pulse with a new time argument:

$$\mathbf{E}(\mathbf{r}_0+\Delta\mathbf{r},t) = \mathbf{E}\left(\mathbf{r}_0,t-t'\right)e^{i\left(\mathbf{k}(\omega_0)\cdot\Delta\mathbf{r}-\omega_0 t'\right)} \tag{7.36}$$

Notice that (7.32) for propagating in vacuum agrees with this result, since $\mathbf{k}_{\text{vac}}(\omega_0)\cdot\Delta\mathbf{r} = \omega_0 z/c$. The second factor in (7.36) merely gives a phase shift governed by the *phase velocity* of the carrier frequency (see (7.9)):

$$v_p(\omega_0) = \frac{\omega_0}{k(\omega_0)} \tag{7.37}$$

The phase shift vanishes for propagation in vacuum. Ignoring the phase shift, (7.36) is only altered by a delay t', the time required for the pulse to traverse the displacement $\Delta\mathbf{r}$.

The function $\partial\mathbf{k}/\partial\omega\cdot\Delta\mathbf{r}$ is known as the *group delay function*, and in (7.35) it is evaluated at the carrier frequency ω_0. Traditional *group velocity* is obtained by dividing the displacement $\Delta\mathbf{r}$ by the group delay time t' to obtain

$$v_g^{-1}(\omega_0) = \left.\frac{\partial k(\omega)}{\partial\omega}\right|_{\omega_0} \tag{7.38}$$

Group delay (or group velocity) essentially tracks the center of the pulse.

In our derivation we have assumed that the phase delay $\mathbf{k}(\omega)\cdot\Delta\mathbf{r}$ could be well-represented by the first two terms of the expansion (7.33). While this assumption gives results that are often useful, higher-order terms can also play a role. In section 7.5 we'll find that the next term in the expansion controls the rate at which the pulse spreads as it travels. We should also note that there are times when the expansion (7.33) fails to converge (when ω_0 is near a resonance of the medium), and the above expansion approach is not valid. We'll analyze pulse propagation in this sticky situation in section 7.6.

7.5 Quadratic Dispersion

A light pulse traversing a material in general undergoes *dispersion* when different frequency components propagate with different phase velocities. As an example, consider a short laser pulse traversing an optical component such as a lens or window, as depicted in Fig. 7.8. The short light pulse can broaden in time[8] with the different frequency components becoming separated (often called *stretching* or *chirping*). If absorption (and surface reflections) can be neglected, then the amplitude of $\mathbf{E}(\mathbf{r},\omega)$ does not change – only its phase changes – and the power spectrum (7.21) remains unaltered.

Continuing our example of a short pulse traversing a piece of glass, we assume that the pulse travels in the $\hat{\mathbf{z}}$-direction. We place $\mathbf{r}_0$ at the start of the glass where we assign $z=0$, so that $\mathbf{k}\cdot\Delta\mathbf{r}=kz$. We take the Fourier transform of pulse at $z=0$ to determine the amplitudes and phases of the plane waves involved.

To find the waveform at the new position z (where the pulse presumably has just exited the glass), we must adjust the phase of each plane wave by the factor kz and take the inverse Fourier transform accruing to (7.31). Again, the function $k(\omega)$ must be specified. Typically, the functional form of $n(\omega)$ spoils any chance of doing the integral analytically. And as before, we will resort to the expansion (7.33), but this time we will keep an additional term:

$$k(\omega)\,z \cong k_0 z + v_g^{-1}(\omega-\omega_0)\,z + \alpha\,(\omega-\omega_0)^2 z + \cdots \tag{7.39}$$

where

$$k_0 \equiv k(\omega_0) = \frac{\omega_0 n(\omega_0)}{c} \tag{7.40}$$

$$v_g^{-1} \equiv \left.\frac{\partial k}{\partial \omega}\right|_{\omega_0} = \frac{n(\omega_0)}{c} + \frac{\omega_0 n'(\omega_0)}{c} \tag{7.41}$$

$$\alpha \equiv \frac{1}{2}\left.\frac{\partial^2 k}{\partial \omega^2}\right|_{\omega_0} = \frac{n'(\omega_0)}{c} + \frac{\omega_0 n''(\omega_0)}{2c} \tag{7.42}$$

Unfortunately, even after resorting to the expansion we won't be able to perform the inverse Fourier transform except for very specific initial pulses. However, we can get an idea for how quadratic dispersion works by considering the specific example of a Gaussian pulse.

---

[8] See J. D. Jackson, *Classical Electrodynamics*, 3rd ed., Sect. 7.9 (New York: John Wiley, 1999).

Example 7.7

A Gaussian waveform similar to that in Example 7.6 propagates through a piece of glass with thickness $\Delta r = z$. Compute the waveform exiting the glass.

Solution: Again, the Fourier transform of the Gaussian pulse before propagation is given by (7.25):

$$\mathbf{E}(0,\omega) = T\mathbf{E}_0 e^{-\frac{T^2(\omega-\omega_0)^2}{2}}$$

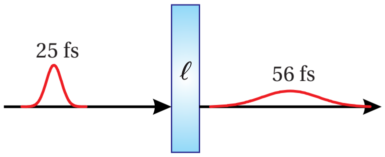


Figure 7.8 A 25 fs pulse traversing an $\ell = 1$ cm piece of BK7 glass.

With the aid of expansion (7.39), the inverse Fourier transform (7.31) (which yields the pulse after propagation) becomes

$$\begin{aligned}\mathbf{E}(z,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathbf{E}_0 T e^{-\frac{T^2(\omega-\omega_0)^2}{2}} e^{ik_0 z + iv_g^{-1}(\omega-\omega_0)z + i\alpha(\omega-\omega_0)^2 z} e^{-i\omega t}\, d\omega \\ &= \frac{T\mathbf{E}_0 e^{i(k_0 z-\omega_0 t)}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(T^2/2 - i\alpha z)(\omega-\omega_0)^2} e^{iv_g^{-1}(\omega-\omega_0)z - i(\omega-\omega_0)t}\, d\omega\end{aligned} \tag{7.43}$$

We can avoid considerable clutter if we change variables to $\omega' \equiv \omega - \omega_0$. Then the inverse Fourier transform becomes

$$\mathbf{E}(z,t) = \frac{T\mathbf{E}_0 e^{i(k_0 z-\omega_0 t)}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{T^2}{2}(1-i2\alpha z/T^2)\omega'^2 - i(t-z/v_g)\omega'}\, d\omega' \tag{7.44}$$

The above integral can be performed with the aid of (0.55). The result is

$$\begin{aligned}\mathbf{E}(z,t) &= \frac{T\mathbf{E}_0 e^{i(k_0 z-\omega_0 t)}}{\sqrt{2\pi}}\sqrt{\frac{\pi}{\frac{T^2}{2}\left(1-i2\alpha z/T^2\right)}}\, e^{-\frac{\left(t-z/v_g\right)^2}{4\frac{T^2}{2}\left(1-i2\alpha z/T^2\right)}} \\ &= \mathbf{E}_0 e^{i(k_0 z-\omega_0 t)}\frac{e^{\frac{i}{2}\tan^{-1}\frac{2\alpha z}{T^2}}}{\sqrt[4]{1+\left(2\alpha z/T^2\right)^2}}\, e^{-\frac{\left(t-z/v_g\right)^2}{2T^2\left(1+\left(2\alpha z/T^2\right)^2\right)}\left(1+i2\alpha z/T^2\right)}\end{aligned} \tag{7.45}$$

Next, we spruce up the appearance of this rather cumbersome formula as follows:

$$\mathbf{E}(z,t) = \frac{\mathbf{E}_0}{\sqrt{\tilde{T}(z)/T}}\, e^{-\frac{\left(t-z/v_g\right)^2}{2\tilde{T}^2(z)}}\, e^{-i\frac{\left(t-z/v_g\right)^2}{2\tilde{T}^2(z)}\Phi(z) + i(k_0 z-\omega_0 t) + i\frac{1}{2}\tan^{-1}\Phi(z)} \tag{7.46}$$

where

$$\Phi(z) \equiv \frac{2\alpha}{T^2} z \tag{7.47}$$

and

$$\tilde{T}(z) \equiv T\sqrt{1+\Phi^2(z)} \tag{7.48}$$

We can immediately make a few observations about (7.46). First, note that at $z = 0$ (i.e. zero thickness of glass), (7.46) reduces to the input pulse $\mathbf{E}(0,t) = \mathbf{E}_0 e^{-t^2/2T^2} e^{-i\omega_0 t}$, as it should. Secondly, the peak of the pulse moves at speed v_g since the factor $e^{-(t-z/v_g)^2/2\tilde{T}^2(z)}$ controls the pulse amplitude, while the other terms (multiplied by i) in the exponent of (7.46) merely alter the phase. Also note that the duration of the pulse increases and its peak intensity decreases as it travels, since $\tilde{T}(z)$ increases with z. In P7.8 we will find that (7.46) also predicts that for large z, the field of the spread-out pulse oscillates less rapidly at the beginning of the pulse than at the end (assuming $\alpha > 0$). This phenomenon, known as pulse *chirping*, means that red frequencies get ahead of blue frequencies during propagation since the red frequencies experience a lower index of refraction.

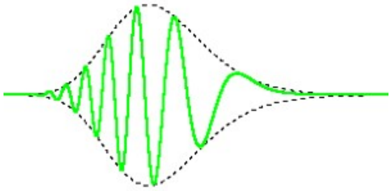


Figure 7.9 Animation of a Gaussian-envelope pulse (electric field) undergoing dispersion during transit.

While Example 7.7 is worked out for the specific case of a Gaussian pulse, the results are qualitatively similar for all pulses. The exact details vary with pulse shape, but all short pulses eventually broaden and chirp as they propagate through a dispersive medium such as glass. Higher-order terms in the expansion (7.33) that were neglected cause additional spreading, chirping, and other deformations to the pulses as they propagate. The influence of each order becomes progressively more cumbersome to study analytically. It is easier to perform the inverse Fourier transform numerically; there is no need to resort to the expansion of $\mathbf{k}(\omega)$ if the integration is done numerically.

7.6 Generalized Context for Group Delay

The expansion of $\mathbf{k}(\omega)$ in (7.33) is inconvenient if the frequency content (bandwidth) of a waveform encompasses a substantial portion of a resonance structure. In this case, it becomes necessary to retain a large number of terms in (7.33) to describe accurately the phase delay $\mathbf{k}(\omega)\cdot\Delta\mathbf{r}$. Moreover, if the bandwidth of the waveform is wider than the spectral resonance of the medium, the series altogether fails to converge. These difficulties have led to the traditional viewpoint that group velocity loses meaning for broadband waveforms near a resonance. In this section, we study a broader context for group velocity (or rather its inverse, group delay $dk/d\omega$), which is always valid, even for broadband pulses where the expansion (7.33) utterly fails. The analysis avoids the expansion and so is not restricted to a narrowband context. Since the imaginary part of the index becomes important near a resonance, we will need to treat k as complex.

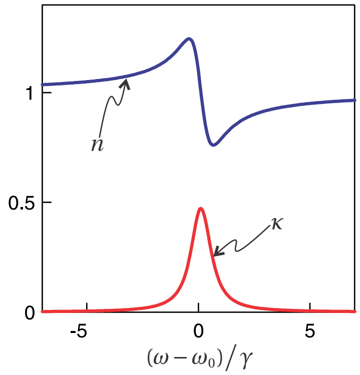


Figure 7.10 Real and imaginary parts of the refractive index for an absorptive medium.

We are interested in the *arrival time* of a waveform (or pulse) to a point, say, where a detector is located. The definition of the arrival time of pulse energy need only involve the Poynting flux (or the intensity), since it alone is responsible for energy transport. To deal with arbitrary broadband pulses, the arrival time should avoid presupposing a specific pulse shape, since the pulse may evolve in complicated ways during propagation. For example, the pulse peak or the midpoint on the rising edge of a pulse are poor indicators of arrival time if the pulse contains multiple peaks or a long and nonuniform rise time.

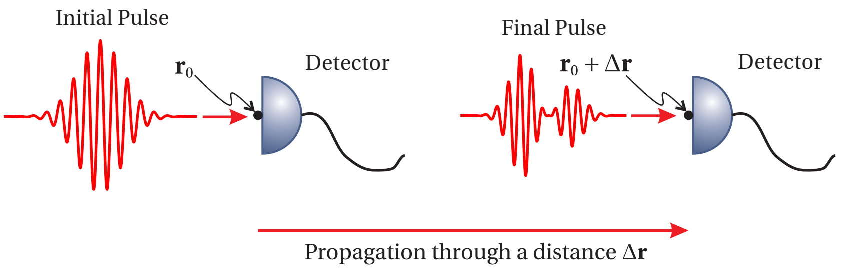

Initial Pulse
$\mathbf{r}_0$
Detector
Final Pulse
$\mathbf{r}_0 + \Delta\mathbf{r}$
Detector
Propagation through a distance $\Delta\mathbf{r}$

Figure 7.11 Transit time defined as the difference between arrival time at two points.

For the reasons given, we use a time expectation integral (or time 'center-of-mass') to describe the arrival time of a pulse:

$$\langle t \rangle_{\mathbf{r}} \equiv \frac{\int_{-\infty}^{\infty} t I(\mathbf{r}, t)\, dt}{\int_{-\infty}^{\infty} I(\mathbf{r}, t)\, dt} \tag{7.49}$$

For simplification, we have assumed that the light travels in a uniform direction by using intensity rather than the Poynting vector.

Consider a pulse as it travels from point $\mathbf{r}_0$ to point $\mathbf{r} = \mathbf{r}_0 + \Delta\mathbf{r}$ in a homogeneous medium. The difference in arrival times at the two points is

$$\Delta t \equiv \langle t \rangle_{\mathbf{r}} - \langle t \rangle_{\mathbf{r}_0} \tag{7.50}$$

The pulse shape can evolve in complicated ways between the two points, spreading with different portions being absorbed (or amplified) during transit as depicted in Fig. 7.11. Nevertheless, (7.50) renders an unambiguous time interval between the passage of the pulse center at each point.

This difference in arrival time can be shown to consist of two terms (see P7.11):[9]

$$\Delta t = \Delta t_G(\mathbf{r}) + \Delta t_R(\mathbf{r}_0) \tag{7.51}$$

The first term, called the *net group delay*, dominates if the field waveform is initially symmetric in time (e.g. an unchirped Gaussian). It amounts to a spectral average of the group delay function taken with respect to the spectral content of the pulse arriving at the final point $\mathbf{r} = \mathbf{r}_0 + \Delta\mathbf{r}$:

$$\Delta t_G(\mathbf{r}) = \frac{\int_{-\infty}^{\infty} I(\mathbf{r}, \omega) \left(\frac{\partial \mathrm{Re}\,\mathbf{k}}{\partial \omega} \cdot \Delta\mathbf{r} \right) d\omega}{\int_{-\infty}^{\infty} I(\mathbf{r}, \omega)\, d\omega} \tag{7.52}$$

where $I(\mathbf{r}, \omega)$ is given in (7.21). The two curves in Fig. 7.12 show $I(\mathbf{r}_0, \omega)$ (before propagation) and $I(\mathbf{r}, \omega)$ (after propagation) for an initially Gaussian pulse. As

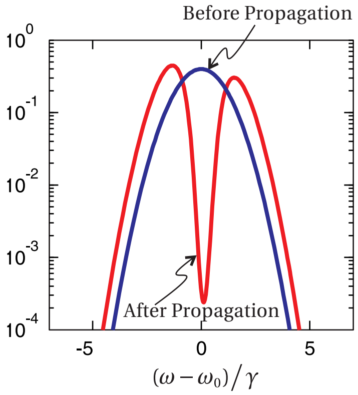


Figure 7.12 *Normalized* power spectrum of a broadband pulse before and after propagation through an absorbing medium with the complex index shown in Fig. 7.10. The absorption line eats a hole in the spectrum.

---

[9] M. Ware, S. A. Glasgow, and J. Peatross, "The Role of Group Velocity in Tracking Field Energy in Linear Dielectrics," Opt. Express **9**, 506-518 (2001).

seen in (7.52), the pulse travel time depends on the spectral shape of the pulse at the end of propagation.

Note the close resemblance between the formulas (7.49) and (7.52). Both are expectation integrals. The former is executed as a 'center-of-mass' integral on time; the latter is executed in the frequency domain on $\partial \mathrm{Re}\mathbf{k}\cdot\Delta\mathbf{r}/\partial\omega$, the group delay function (7.38). The group delay at every frequency present in the pulse influences the result. If the pulse has a narrow bandwidth in the neighborhood of ω_0, the integral reduces to $\partial \mathrm{Re}\mathbf{k}/\partial\omega|_{\omega_0}\cdot\Delta\mathbf{r}$, in agreement with (7.38) (see P7.9). The net group delay depends only on the spectral content of the pulse, independent of its temporal organization (i.e. the phase of $\mathbf{E}(\mathbf{r},\omega)$ has no influence). Only the real part of the $\mathbf{k}$-vector plays a direct role in (7.52).

The second term in (7.51) is the *reshaping delay* Δt_R. It represents a delay that arises solely from a reshaping of the spectral amplitude. Often this term is negligible. The term takes into account how the pulse time center-of-mass shifts as portions of the spectrum are removed (or added), as illustrated in Fig. 7.13. It is computed at $\mathbf{r}_0$ *before propagation takes place*:[10]

$$\Delta t_R(\mathbf{r}_0) = \langle t\rangle_{\mathbf{r}_0}\Big|_{\text{altered}} - \langle t\rangle_{\mathbf{r}_0} \tag{7.53}$$

Here $\langle t\rangle_{\mathbf{r}_0}$ represents the usual arrival time of the pulse at the initial point $\mathbf{r}_0$, according to (7.49). The intensity at this point is associated with a field $\mathbf{E}(\mathbf{r}_0,t)$ whose spectrum is $\mathbf{E}(\mathbf{r}_0,\omega)$. On the other hand, $\langle t\rangle_{\mathbf{r}_0}\big|_{\text{altered}}$ is the arrival time of a pulse with modified spectrum $\mathbf{E}(\mathbf{r}_0,\omega)\,e^{-\mathrm{Im}\mathbf{k}\cdot\Delta\mathbf{r}}$. Notice that $\mathbf{E}(\mathbf{r}_0,\omega)\,e^{-\mathrm{Im}\mathbf{k}\cdot\Delta\mathbf{r}}$ is still evaluated at the initial point $\mathbf{r}_0$. Only the spectral amplitude (not the phase) is modified, according to what is anticipated to be lost (or gained) during the trip. In contrast to the net group delay, the reshaping delay is sensitive to how a pulse is organized. The reshaping delay is negligible if the pulse is initially symmetric (in amplitude and phase) before propagation. The reshaping delay also goes to zero in the narrowband limit, and the total delay reduces to the net group delay.

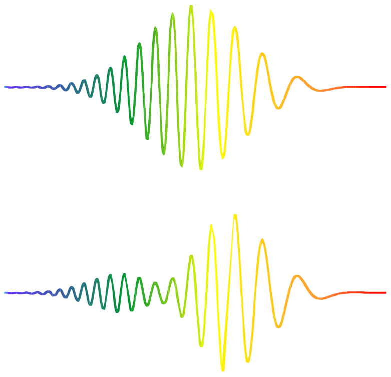


Figure 7.13 The center of a chirped pulse can shift owing to the reshaping effect when a portion of the spectrum is removed.

Example 7.8

Find the time required for a Gaussian pulse (7.23) to traverse a slab of absorption material (neglecting possible surface reflections). Let the material response be described by the Lorentz model described in section 2.3 with the carrier frequency of the pulse ω_0, coinciding with the material resonance frequency. Let the slab have thickness $\Delta r = c\gamma^{-1}/10$ and absorption strength $\omega_p^2 = 10\gamma$.

Solution: The spectrum of the initially Gaussian pulse is given by (7.25), and its power spectrum is[11]

$$I(\mathbf{r}_0,\omega) \propto e^{-T^2(\omega-\omega_0)^2}$$

---

[10] The reshaping delay can instead be computed after propagation takes place, in which case the net group delay should be computed with the initial rather than final spectrum.

[11] In general, one should write $\tilde{\omega}_0$ to distinguish the carrier frequency of the pulse from the resonance frequency of the material ω_0; in practice, these are often different.

After propagating from $\mathbf{r}_0$ to $\mathbf{r} = \mathbf{r}_0 + \Delta r$, the power spectrum becomes

$$I(\mathbf{r},\omega) \propto e^{-T^2(\omega-\omega_0)^2} e^{-2\frac{\kappa(\omega)\omega}{c}\Delta r}$$

The net group delay is then

$$\Delta t_G(\mathbf{r}) = \Delta r \frac{\int_{-\infty}^{\infty} I(\mathbf{r},\omega)\left(\frac{\partial(\omega n/c)}{\partial\omega}\right)d\omega}{\int_{-\infty}^{\infty} I(\mathbf{r},\omega)\,d\omega} = \frac{\Delta r}{c}\frac{\int_{-\infty}^{\infty} e^{-T^2(\omega-\omega_0)^2} e^{-2\frac{\kappa\omega}{c}\Delta r}\left(n+\omega\frac{\partial n}{\partial\omega}\right)d\omega}{\int_{-\infty}^{\infty} e^{-T^2(\omega-\omega_0)^2} e^{-2\frac{\kappa\omega}{c}\Delta r}\,d\omega}$$

The index of refraction $n + i\kappa$ is given by (2.39) (see also (2.27) and (2.29)). Since the expressions for n and κ are complicated, the integration in the above formula must be performed numerically.

The result when $T = T_1 = 10\gamma^{-1}/\sqrt{2}$ (narrowband) is

$$\Delta t_G = -5.1/\gamma = -51\Delta r/c = -0.72T_1$$

and the result when $T = T_2 = \gamma^{-1}/\sqrt{2}$ (broadband) is

$$\Delta t_G = 0.67/\gamma = 6.7\Delta r/c = 0.95T_2$$

The reshaping delay (7.53) in both cases is negligible.

The narrowband pulse (with duration T_1) in Example 7.8 traverses the absorbing medium *superluminally* (i.e. faster than c). The negative transit time means that the 'center-of-mass' of the exiting pulse emerges even before the 'center-of-mass' of the entering pulse reaches the medium! On the other hand, the broadband pulse (with the shorter duration T_2) has a large positive delay time, indicating that the exiting pulse emerges *subluminally*.

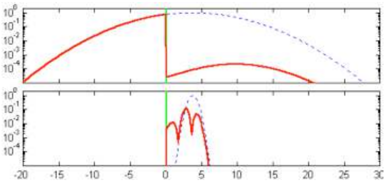


Figure 7.14 Animation comparing narrowband vs. broadband Gaussian pulses traversing an absorbing slab (green stripe) on resonance. Note the logarithmic scale. See Example 7.8.

Figure 7.14 shows the intensity profiles for these two pulses as they traverse the absorption slab, calculated with the aid of (7.31). By eye, one can see how the centers of the two pulses are either advanced or delayed as they go through the absorption medium. In both cases, the pulse that emerges is well within the envelope of the original pulse propagated forward at c. In the case of the broadband pulse, the absorption peak eats a hole in the center of the spectrum as shown in Fig. 7.12, causing the emerging pulse to be distorted in time. The analysis in this section predicts the center of pulses, whereas to see the shape of pulses one needs to calculate (7.31).

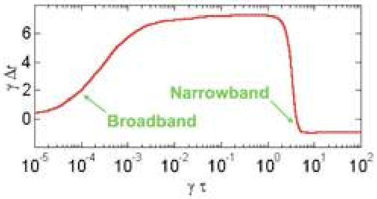


Figure 7.15 Delay as a function of pulse duration.

The results for the two pulse durations in Example 7.8 indicate a trend. Superluminal behavior only occurs for long boring pulses. In the case of a single absorption resonance, this comes with a severe cost of attenuation. Figure 7.15 shows the delay time as a function of pulse duration. As the injected pulse becomes more sharply defined in time, the superluminal behavior does not persist. Sharply defined waveforms (i.e. broadband) cannot propagate superluminally precisely because much of their bandwidth lies away from the frequencies with superluminal group delays.

We should mention that superluminal propagation cannot persist for indefinite distances since the medium eventually removes the superluminal spectral components through absorption (or else adds subluminal spectral components in the case of amplification). This limits the amount that a pulse center can be advanced—on the scale of the pulse's own duration.

As we saw for the absorption situation the exiting pulse is tiny and resides well within the original envelope of the pulse propagated forward at speed c, as depicted in Fig. 7.16. Without the absorbing material in place, the signal would be detectable just as early. This statement is also true for amplifying media.[12] Figure 7.17 shows narrowband and broadband pulses traversing an amplifying medium. In this case, superluminal behavior occurs for spectra near by but not on an amplifying resonance. If the pulse is too broadband, its spectrum will be amplified, which adds slower components to the overall group delay.

While it may be surprising at first to realize that group velocity can become superluminal, it is to be expected for pulses whose spectra lie in the vicinity of a medium resonance. Group velocity v_g tracks the *presence* of field energy, whether that energy propagates or is extracted from the medium at a point downstream. Energy is never transported faster than the universal speed limit c. A detailed analysis of energy flow is given in Appendix 7.B.

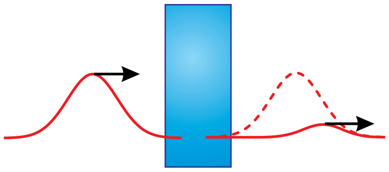


Figure 7.16 Narrowband pulse traversing an absorbing medium.

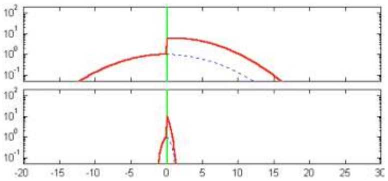


Figure 7.17 Animation comparing narrowband vs. broadband Gaussian pulses traversing an amplifying slab (green stripe) slightly off resonance.

Appendix 7.A Pulse Chirping in a Grating Pair

Reflection grating pairs can be used to introduce large amounts of dispersion into a light pulse. Gratings are especially useful for amplification of ultrashort laser pulses, where laser pulses are first stretched in time before amplification (to prevent damage to the amplifier) and then compressed back to short duration just before the experiment (called *chirped pulse amplification*). Diffraction from a grating causes each **k**-vector to travel at a different angle. A second grating parallel to the first can realign all of the **k**-vectors to be parallel to each other. Since laser beams are not infinitely wide, the light is typically sent through the grating pair twice to undo the tendency of the different frequency components to become laterally separated. In the present analysis, we will consider an infinitely wide plane wave pulse incident upon a grating. The scenario is depicted in Fig. 7.18: A short plane wave pulse strikes the grating at an angle, and a spreading pulse emerges.

Consider a plane-wave pulse that ricochets between a pair of parallel grating surfaces. Although different **k**-vectors point with different angles, they are all straightened out upon diffracting from the second grating. For simplicity, we will consider a pulse just before the first bounce and just after the second bounce, even though we are interested in the dispersion that takes place between the gratings. This allows us to treat the **k**-vectors as being parallel for purposes of computing intensity.

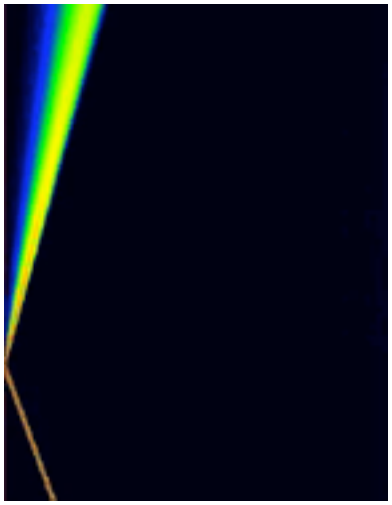


Figure 7.18 Animation showing a short plane-wave pulse diffracting from a grating positioned along the left edge of the frame.

---

[12] You can use the Lorentz model (2.40) to describe an amplifying medium with a negative oscillator strength f.

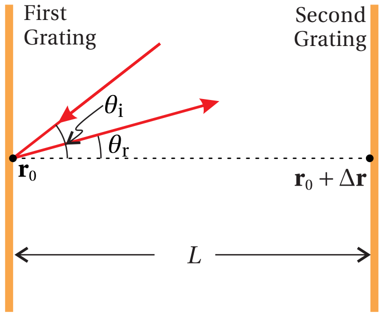


Figure 7.19 Direction of **k**-vector between parallel gratings (top view). Grating rulings run in and out of the page.

Consider a plane wave incident on a grating at an incident angle θ_i with respect to the grating normal (aligned with the x-axis in our coordinate system) as depicted in Fig. 7.19. The plane wave diffracts from the first grating at an angle θ_r (also referenced from the grating normal). This angle is governed by the grating diffraction formula[13]

$$\theta_\mathrm{r}(\omega) = \sin^{-1}\left(\frac{2\pi c}{\omega d} - \sin\theta_i\right) \tag{7.54}$$

where d is the grating groove spacing. By examining the geometry of the figure, we see that the reflected **k**-vector is given by $\mathbf{k} = \left(\hat{\mathbf{x}}\cos\theta_\mathrm{r} + \hat{\mathbf{y}}\sin\theta_\mathrm{r}\right)\omega/c$.

Suppose we know the pulse at a point $\mathbf{r}_0$ on the first grating. Next we choose a point $\mathbf{r}_0 + \Delta\mathbf{r}$ on the second grating where we will determine the outgoing pulse. Since we are considering an infinitely wide plane-wave pulse, it doesn't matter where we choose that point as long as it lies on the surface of the second grating. The waveform will be the same everywhere on the second grating, only with a different arrival time. For convenience, we might as well take the second point to be $\mathbf{r}_0 + \Delta\mathbf{r} = \mathbf{r}_0 + L\hat{\mathbf{x}}$ as shown in Fig. 7.19.

The phase delay needed for (7.30) becomes

$$\mathbf{k}(\omega)\cdot\Delta\mathbf{r} = \frac{L\omega}{c}\cos\theta_\mathrm{r} \tag{7.55}$$

We will express this as a Taylor-series expansion similar to (7.39) so that we can perform the inverse Fourier transform analytically. We will approximate (7.55) as

$$\mathbf{k}(\omega)\cdot\Delta\mathbf{r} \approx k_0 L + v_g^{-1}(\omega-\omega_0)L + \alpha(\omega-\omega_0)^2 L + \cdots \tag{7.56}$$

so that we can take advantage of formula (7.46). To calculate the terms in this expansion we will need the derivative of (7.54):

$$\begin{aligned}\frac{d\theta_\mathrm{r}}{d\omega} &= \frac{1}{\sqrt{1-\left(\frac{2\pi c}{\omega d}-\sin\theta_\mathrm{i}\right)^2}}\left(-\frac{2\pi c}{\omega^2 d}\right) = \frac{1}{\sqrt{1-\sin^2\theta_\mathrm{r}}}\left(-\frac{2\pi c}{\omega^2 d}\right)\\ &= -\frac{2\pi c}{\omega^2 d\cos\theta_\mathrm{r}} = -\frac{\sin\theta_\mathrm{i}+\sin\theta_\mathrm{r}}{\omega\cos\theta_\mathrm{r}}\end{aligned} \tag{7.57}$$

The derivatives of (7.55) necessary for the Taylor's series expansion are

$$\begin{aligned}\frac{d\mathbf{k}}{d\omega}\cdot\Delta\mathbf{r} &= \frac{L}{c}\left(\cos\theta_\mathrm{r} - \omega\sin\theta_\mathrm{r}\frac{d\theta_\mathrm{r}}{d\omega}\right)\\ &= \frac{L}{c}\left(\cos\theta_\mathrm{r} + \sin\theta_\mathrm{r}\frac{\sin\theta_\mathrm{i}+\sin\theta_\mathrm{r}}{\cos\theta_\mathrm{r}}\right)\\ &= \frac{L}{c}\left(\frac{1+\sin\theta_\mathrm{r}\sin\theta_\mathrm{i}}{\cos\theta_\mathrm{r}}\right)\end{aligned} \tag{7.58}$$

and

---

[13] This formula is equivalent to $d\sin\theta_\mathrm{i} + d\sin\theta_\mathrm{r} = \lambda$ with $\lambda = 2\pi c/\omega$.

$$\begin{aligned}\frac{d^2\mathbf{k}}{d\omega^2}\cdot\Delta\mathbf{r} &= \frac{L}{c}\left(\sin\theta_\mathrm{i}+\frac{\sin\theta_\mathrm{r}\,(1+\sin\theta_\mathrm{r}\sin\theta_\mathrm{i})}{\cos^2\theta_\mathrm{r}}\right)\frac{d\theta_\mathrm{r}}{d\omega}\\ &= \frac{L}{c}\left(\frac{\sin\theta_\mathrm{i}+\sin\theta_\mathrm{r}}{\cos^2\theta_\mathrm{r}}\right)\left(-\frac{\sin\theta_\mathrm{i}+\sin\theta_\mathrm{r}}{\omega\cos\theta_\mathrm{r}}\right)\\ &= -\frac{L}{\omega c}\frac{(\sin\theta_\mathrm{i}+\sin\theta_\mathrm{r})^2}{\cos^3\theta_\mathrm{r}}\end{aligned} \tag{7.59}$$

The coefficients in (7.56) then are

$$k_0 \equiv \mathbf{k}|_{\omega_0}\cdot\frac{\Delta\mathbf{r}}{L} = \frac{\omega_0}{c} \tag{7.60}$$

$$v_g^{-1} \equiv \left.\frac{d\mathbf{k}}{d\omega}\right|_{\omega_0}\cdot\frac{\Delta\mathbf{r}}{L} = \left.\frac{1+\sin\theta_\mathrm{r}\sin\theta_\mathrm{i}}{c\cos\theta_\mathrm{r}}\right|_{\omega_0} \tag{7.61}$$

$$\alpha \equiv \frac{1}{2}\left.\frac{d^2\mathbf{k}}{d\omega^2}\right|_{\omega_0}\cdot\frac{\Delta\mathbf{r}}{L} = -\left.\frac{(\sin\theta_\mathrm{i}+\sin\theta_\mathrm{r})^2}{2c\omega\cos^3\theta_\mathrm{r}}\right|_{\omega_0} \tag{7.62}$$

In the case of a Gaussian pulse, we can employ (7.46), where L takes the place of z, and k_0, v_g^{-1} and α are defined by (7.60) – (7.62). The duration of the pulse is controlled by (7.62) and the spacing between the gratings L.

Appendix 7.B Causality and Exchange of Energy with the Medium

As shown in section 7.6, the group delay function is useful for predicting when the centroid of a light pulse will arrive to a point in space. Since this is only part of the whole energy story, there is no problem when it becomes superluminal. The overly rapid appearance of electromagnetic energy at one point and its simultaneous disappearance at another point merely indicates an exchange of energy between the electric field and the medium.[14]

We need not be dazzled by a magician who invites the audience to look only at the field energy while energy transfers into and out of the 'unwatched' domain of the medium. Extra field energy seems to appear 'prematurely' downstream only if there is already nonzero field energy downstream to stimulate a transfer of energy from the medium. The actual transport of energy is strictly bounded by c; superluminal propagation of a sharp *signal front* is impossible.

In accordance with Poynting's theorem (2.51), the total energy density stored in an electromagnetic field and in a medium is given by

$$u(\mathbf{r},t) = u_\text{field}(\mathbf{r},t) + u_\text{med}(\mathbf{r},t) + u(\mathbf{r},-\infty) \tag{7.63}$$

---

[14] M. Ware, S. A. Glasgow, and J. Peatross, "Energy Transport in Linear Dielectrics," Opt. Express **9**, 519-532 (2001).

where the time-dependent accumulation of energy transferred into the medium from the field (ignoring possible free current $\mathbf{J}_{\text{free}}$) is

$$u_{\text{med}}(\mathbf{r},t) = \int_{-\infty}^{t} \mathbf{E}\left(\mathbf{r},t'\right)\cdot\frac{\partial\mathbf{P}\left(\mathbf{r},t'\right)}{\partial t'}\,dt' \tag{7.64}$$

The expression (7.63) for the energy density includes all (relevant) forms of energy, including a nonzero integration constant $u(\mathbf{r},-\infty)$ corresponding to energy stored in the medium before the arrival of any pulse (important in the case of an amplifying medium). $u_{\text{field}}(\mathbf{r},t)$ and $u_{\text{med}}(\mathbf{r},t)$ are both zero before the arrival of the pulse (i.e. at $t=-\infty$). In addition, $u_{\text{field}}(\mathbf{r},t)$, given by (2.53), returns to zero after the pulse has passed (i.e. at $t=+\infty$).

As u_{med} increases, the energy in the medium increases. Conversely, as u_{med} decreases, the medium surrenders energy to the electromagnetic field. While it is possible for u_{med} to become negative, the combination $u_{\text{med}}+u(-\infty)$ (i.e. the net energy in the medium) can never go negative since a material cannot surrender more energy than it possesses to begin with.

Poynting's theorem (2.51) has the form of a continuity equation which when integrated spatially over a small volume V yields

$$\oint_A \mathbf{S}\cdot d\mathbf{a} = -\frac{\partial}{\partial t}\int_V u\,dV \tag{7.65}$$

where the left-hand side has been transformed into an surface integral (via the divergence theorem (0.11)) representing the power leaving the volume. Let the volume be small enough to take $\mathbf{S}$ to be uniform throughout V.

We can define an *energy transport velocity* (directed along $\mathbf{S}$) as the *effective* speed at which all of the energy density would need to travel in order to achieve the Poynting flux:

$$\mathbf{v}_E \equiv \frac{\mathbf{S}}{u} \tag{7.66}$$

Note that this ratio of the Poynting flux to the energy density has units of velocity. When the total energy density u is used in computing (7.66), the energy transport velocity has a *fictitious* nature; it is not the actual velocity of the total energy (since part is stationary), but rather the effective velocity necessary to achieve the same energy transport that the electromagnetic flux alone delivers. If we reduce the denominator to the subset of the energy that can move, namely u_{field}, the Cauchy-Schwartz inequality (i.e. $\alpha^2+\beta^2\geq 2\alpha\beta$) ensures an energy transport velocity v_E remains strictly bounded by the speed of light in vacuum c. The total energy density u is at least as great as the field energy density u_{field}. Hence, this strict luminality is maintained.

Centroid of Energy

Consider a weighted average of the energy transport velocity:

$$\langle \mathbf{v}_E \rangle \equiv \frac{\int \mathbf{v}_E u \, d^3 r}{\int u \, d^3 r} = \frac{\int \mathbf{S} \, d^3 r}{\int u \, d^3 r} \tag{7.67}$$

where we have substituted from (7.66).

Integration by parts leads to

$$\langle \mathbf{v}_E \rangle = -\frac{\int \mathbf{r} \nabla \cdot \mathbf{S} \, d^3 r}{\int u \, d^3 r} = \frac{\int \mathbf{r} \frac{\partial u}{\partial t} \, d^3 r}{\int u \, d^3 r} \tag{7.68}$$

where we have assumed that the volume for the integration encloses *all* energy in the system and that the field near the edges of this volume is zero. Since we have included all energy, Poynting's theorem (2.51) can be written with no source terms (i.e. $\nabla \cdot \mathbf{S} + \partial u / \partial t = 0$). This means that the total energy in the system is conserved and is given by the integral in the denominator of (7.68). This allows the derivative to be brought out in front of the entire expression giving

$$\langle \mathbf{v}_E \rangle = \frac{\partial \langle \mathbf{r} \rangle}{\partial t} \quad \text{where} \quad \langle \mathbf{r} \rangle \equiv \frac{\int \mathbf{r} u \, d^3 r}{\int u \, d^3 r} \tag{7.69}$$

The latter expression represents the 'center-of-mass' or *centroid* of the total energy in the system, which is guaranteed to evolve strictly luminally since $\mathbf{v}_E$ is everywhere luminal.[15]

It is enlightening to consider u_{med} within a frequency-domain context. In an isotropic medium, the polarization for an individual plane wave can be written in terms of the linear susceptibility defined in (2.16):

$$\mathbf{P}(\mathbf{r}, \omega) = \epsilon_0 \chi(\mathbf{r}, \omega) \mathbf{E}(\mathbf{r}, \omega) \tag{7.70}$$

We can use this to express u_{med} in terms of the electric field and material susceptibility.

Scott A. Glasgow (1964–, American) was born in Santa Fe, New Mexico. While Scott Glasgow does not have an entry in Wikipedia, the authors of this book think he is a great guy. Prof. Glasgow teaches mathematics at Brigham Young University. He worked out the analysis presented in this appendix, including the fact that a linear medium responds to the instantaneous spectrum, which explains within the framework of a spectral analysis why a medium treats the front and back of a pulse differently. Prof. Glasgow is a competitive weight lifter and the father of five children.

Expressing u_{med} in terms of the power spectrum

---

[15] Although (7.69) guarantees that the centroid of the *total* energy moves strictly luminally, there is no such limitation on the centroid of field energy alone. The steps leading to (7.69) are not possible if u_{field} is used in place of u. Explicitly, that is

$$\left\langle \frac{\mathbf{S}}{u_{\text{field}}} \right\rangle \neq \frac{\partial}{\partial t} \frac{\int \mathbf{r} u_{\text{field}} d^3 r}{\int u_{\text{field}} d^3 r}$$

As was pointed out, the left-hand side is strictly luminal. However, the right-hand side can easily exceed c as the medium exchanges energy with the field. In an amplifying medium, for example, the rapid appearance of a pulse downstream can occur when the leading portion of a pulse stimulates energy already present in the medium to convert to the form of field energy. Group velocity is related to this method of accounting, which is why it also can become superluminal.

The field $\mathbf{E}(\mathbf{r}, t)$ can be expressed as an inverse Fourier transform (7.18). Similarly, the polarization $\mathbf{P}$ can be written as[16]

$$\mathbf{P}(\mathbf{r}, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbf{P}(\mathbf{r}, \omega)\, e^{-i\omega t} d\omega \Rightarrow \frac{\partial \mathbf{P}(\mathbf{r}, t)}{\partial t} = \frac{-i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \omega \mathbf{P}(\mathbf{r}, \omega)\, e^{-i\omega t} d\omega \tag{7.71}$$

The energy density in the medium (7.64) can then be written as

$$u_{\text{med}}(\mathbf{r}, \infty) = \int_{-\infty}^{\infty} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbf{E}\left(\mathbf{r}, \omega'\right) e^{-i\omega' t'} d\omega' \right] \cdot \left[\frac{-i\epsilon_0}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \omega \chi(\mathbf{r}, \omega)\, \mathbf{E}(\mathbf{r}, \omega)\, e^{-i\omega t'} d\omega \right] dt' \tag{7.72}$$

where we have incorporated (7.70) and evaluated u_{med} after the pulse is over at $t = \infty$. We may change the order of integration and write

$$u_{\text{med}}(\mathbf{r}, \infty) = -i\epsilon_0 \int_{-\infty}^{\infty} d\omega \omega \chi(\mathbf{r}, \omega)\, \mathbf{E}(\mathbf{r}, \omega) \cdot \int_{-\infty}^{\infty} d\omega' \mathbf{E}\left(\mathbf{r}, \omega'\right) \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i(\omega+\omega')t'} dt' \tag{7.73}$$

The final integral is a delta function a delta function similar to (0.54), which allows the middle integral also to be performed. The expression for u_{med} then reduces to

$$u_{\text{med}}(\mathbf{r}, \infty) = -i\epsilon_0 \int_{-\infty}^{\infty} \omega \chi(\mathbf{r}, \omega)\, \mathbf{E}(\mathbf{r}, \omega) \cdot \mathbf{E}(\mathbf{r}, -\omega)\; d\omega \tag{7.74}$$

In this derivation, we take $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{P}(\mathbf{r}, t)$ to be real functions, so we can employ the symmetry (7.29) along with

$$\mathbf{P}^*(\mathbf{r}, \omega) = \mathbf{P}(\mathbf{r}, -\omega) \quad \text{and} \quad \chi^*(\mathbf{r}, \omega) = \chi(\mathbf{r}, -\omega).$$

Then we obtain

$$u_{\text{med}}(\mathbf{r}, \infty) = \epsilon_0 \int_{-\infty}^{\infty} \omega \operatorname{Im} \chi(\mathbf{r}, \omega)\, \mathbf{E}(\mathbf{r}, \omega) \cdot \mathbf{E}^*(\mathbf{r}, \omega)\; d\omega \tag{7.75}$$

The expression (7.75) describes the net energy density transferred to a point in the medium after all action has finished (i.e. at $t = \infty$). It involves the power spectrum of the pulse. We can modify this formula in an intuitive way so that it describes the transfer of energy density to the medium for any time during the pulse.

Since the medium is unable to anticipate the spectrum of the entire pulse before experiencing it, the material responds to the pulse according to the history of the field up to each instant. In particular, the material has to be prepared for the possibility of an abrupt cessation of the pulse at any moment, in which case all exchange of energy with the medium immediately ceases. In this extreme scenario, there is no possibility for the medium to recover from previously incorrect attenuation or amplification, so it must have gotten it right already.

---

[16] We assume that the real forms of the fields in the time domain are used for the sake of this multiplication.

If the pulse were in fact to abruptly terminate at a given instant, it would not be necessary to integrate the inverse Fourier transform (7.19) beyond the termination time t after which all contributions are zero. *Causality* requires that the medium be indifferent to whether a pulse actually terminates if that possibility lies in the future. Therefore, (7.75) can apply for any time t (not just for $t = \infty$) if the spectrum (7.19) is evaluated just for that portion of the field previously experienced by the medium (up to time t).

The following is then an exact representation for the energy density (7.64) transferred to the medium:

$$u_{\text{med}}(\mathbf{r}, t) = \epsilon_0 \int_{-\infty}^{\infty} \omega \operatorname{Im}\chi(\mathbf{r}, \omega)\, \mathbf{E}_t(\mathbf{r}, \omega) \cdot \mathbf{E}_t^*(\mathbf{r}, \omega)\; d\omega \tag{7.76}$$

where

$$E_t(\mathbf{r}, \omega) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} E(\mathbf{r}, t')\, e^{i\omega t'}\, dt' \tag{7.77}$$

This time dependence enters only through $\mathbf{E}_t(\mathbf{r}, \omega) \cdot \mathbf{E}_t^*(\mathbf{r}, \omega)$, known as the *instantaneous power spectrum.*

The expression (7.76) gives physical insight into the manner in which causal dielectric materials exchange energy with different parts of an electromagnetic pulse. Since the function $E_t(\omega)$ is the Fourier transform of the pulse truncated at the current time t and set to zero thereafter, it can include many frequency components that are not present in the pulse taken in its entirety. This explains why the medium can respond differently to the front of a pulse compared to the back. Even though absorption or amplification resonances may lie outside of the spectral envelope of a pulse taken in its entirety, the instantaneous spectrum on a portion of the pulse can momentarily lap onto or off of resonances in the medium.

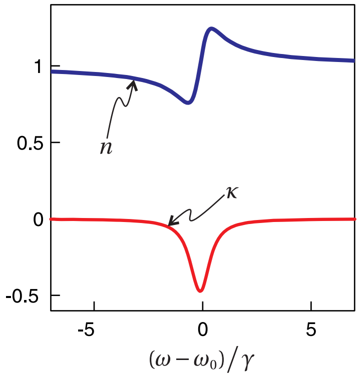


Figure 7.20 Real and imaginary parts of the refractive index for an amplifying medium.

In view of (7.76) and (7.77) it is straightforward to predict when the electromagnetic energy of a pulse will exhibit superluminal or subluminal behavior. In section 7.5, we saw that this behavior is controlled by the group velocity function. However, in (7.76) and (7.77), we see that it is also predictable from the imaginary part of the susceptibility $\chi(\mathbf{r}, \omega)$.

If the entire pulse passing through point $\mathbf{r}$ has a spectrum in the neighborhood of an amplifying resonance, but not on the resonance, superluminal behavior can result. The instantaneous spectrum during the front portion of the pulse is generally wider and can therefore lap onto the nearby gain peak. The medium accordingly amplifies this perceived spectrum, and the front of the pulse grows. The energy is then returned to the medium from the latter portion of the pulse as the instantaneous spectrum narrows and withdraws from the gain peak. The effect is not only consistent with the principle of causality, it is a direct and general consequence of causality as demonstrated by (7.76) and (7.77).

As an illustration, consider the broadband waveform with $T_2 = \gamma^{-1}/\sqrt{2}$ described in Example 7.8. Consider an amplifying medium with index shown in Fig. 7.20 with the amplifying resonance (negative oscillator strength) set on the frequency $\omega_0 = \tilde{\omega}_0 + 2\gamma$, where $\tilde{\omega}_0$ is the carrier frequency. Thus, the resonance structure is centered a modest distance above the carrier frequency, and there is only minor spectral overlap between the pulse and the resonance structure.

Fig. 7.21 shows how the early portion of a pulse has a wide instantaneous spectrum computed by (7.77) that laps onto the amplifying resonance. As the wings grow and access the neighboring resonance, the pulse extracts more energy from the medium. As the wings diminish, the pulse surrenders much of that energy back to the medium, which shifts the center of the pulse forward producing a superluminal effect.

In this appendix we have indirectly proven that a sharply defined signal edge cannot propagate faster than c. If a signal edge begins abruptly at time t_0, the instantaneous spectrum $E_t(\omega)$ clearly remains identically zero until that time. In other words, no energy may be exchanged with the medium until the field energy from the pulse arrives. Since, as was pointed out in connection with (7.66), the Cauchy-Schwartz inequality prevents the field energy from traveling faster than c, at no point in the medium can a signal front exceed c.

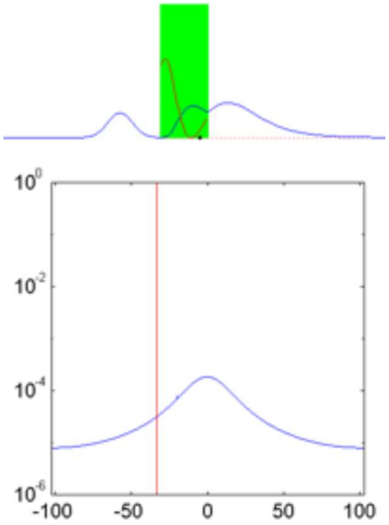


Figure 7.21 Animation of a narrowband pulse traversing an amplifying medium off resonance. The black dot shows the movement of the center of all energy. The red line inside the medium shows the energy held in that medium, which cannot go negative. The lower figure shows the instantaneous spectrum of the pulse at the front of the medium relative to the narrow amplifying resonance.

Appendix 7.C Kramers-Kronig Relations

In the late 1920s, of Hendrik Kramers and Ralph Kronig independently discovered a remarkable relationship between the real and imaginary parts of a material's susceptibility $\chi(\omega)$. Recall that the susceptibility as defined in (2.16) relates the polarization of a material to the field that stimulates the medium:

$$\mathbf{P}(\omega) = \epsilon_0 \chi(\omega) \mathbf{E}(\omega) \tag{7.78}$$

They made an argument based on *causality* (i.e. effect cannot precede cause), which allows one to obtain the real part of $\chi(\omega)$ from the imaginary part of $\chi(\omega)$, if it is known for all ω. Similarly, one can obtain the imaginary part of $\chi(\omega)$ from the real part of $\chi(\omega)$. We develop the Kramers-Kronig formulas below.[17]

We can replace $\mathbf{E}(\omega)$ in (7.78) with the Fourier transform of $\mathbf{E}(t)$ in accordance with (7.19). In addition, we take the inverse Fourier transform (7.19) of both sides of (7.78) and obtain

$$\mathbf{P}(t) = \frac{\epsilon_0}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \chi(\omega) \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathbf{E}(t') e^{i\omega t'} dt' \right] e^{-i\omega t} d\omega \tag{7.79}$$

---

[17] See J. D. Jackson, *Classical Electrodynamics*, 3rd ed., Sect. 7.10 (New York: John Wiley, 1999). Also B. Y.-K. Hu, "Kramers-Kronig in two lines," Am. J. Phys. **57**, 821 (1989).

Next we interchange the order of integration to get

$$\mathbf{P}(t) = \frac{\epsilon_0}{2\pi} \int_{-\infty}^{\infty} \mathbf{E}(t') \left[\int_{-\infty}^{\infty} \chi(\omega) e^{-i\omega(t-t')} d\omega \right] dt' \tag{7.80}$$

Now for the causality argument: The polarization of the medium $\mathbf{P}(t)$ cannot depend on the field $\mathbf{E}(t')$ at future times $t' > t$. Therefore the expression in square brackets must be identically zero unless $t - t' > 0$. This places a restriction on the functional form of $\chi(\omega)$ as we shall see.

The causality argument comes explicitly into play when we employ the following integral formula:[18]

$$e^{-i\omega(t-t')} = \text{sign}\{t - t'\} \frac{1}{i\pi} \int_{-\infty}^{\infty} \frac{e^{-i\omega'(t-t')}}{\omega - \omega'} d\omega' \tag{7.81}$$

Apparently, we require the positive sign since

$$\text{sign}\{t - t'\} \equiv \begin{cases} +1 & (t > t') \\ -1 & (t < t') \end{cases}$$

Upon substitution of (7.81) into (7.80) and after changing the order of integration within the square brackets we obtain

$$\mathbf{P}(t) = \frac{\epsilon_0}{2\pi} \int_{-\infty}^{\infty} \mathbf{E}(t') \left[\int_{-\infty}^{\infty} \left(\frac{1}{i\pi} \int_{-\infty}^{\infty} \frac{\chi(\omega)}{\omega - \omega'} d\omega \right) e^{-i\omega'(t-t')} d\omega' \right] dt' \tag{7.82}$$

For (7.80) and (7.82) to be the same, we require

$$\chi(\omega) = \frac{1}{i\pi} \int_{-\infty}^{\infty} \frac{\chi(\omega')}{\omega' - \omega} d\omega' \tag{7.83}$$

or

$$\text{Re}\chi(\omega) + i\text{Im}\chi(\omega) = \frac{1}{i\pi} \int_{-\infty}^{\infty} \frac{\text{Re}\chi(\omega') + i\text{Im}\chi(\omega')}{\omega' - \omega} d\omega' \tag{7.84}$$

Finally, equating separately the real and imaginary parts of the above equation yields

$$\text{Re}\chi(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\text{Im}\chi(\omega')}{\omega' - \omega} d\omega' \quad \text{and} \quad \text{Im}\chi(\omega) = -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\text{Re}\chi(\omega')}{\omega' - \omega} d\omega' \tag{7.85}$$

---

[18] This integral, which is a specific instance of *Cauchy's theorem*, is tricky because it involves two diverging pieces, to either side of the singularity $\omega = \omega'$. The divergences have opposite sign so that they cancel. The integration must approach the singularity in the same manner from either side, in which case the result is called the *principal value*. In practical terms, if the integral is performed numerically, the sampling of points should straddle the singularity symmetrically; other sampling schemes can change the result dramatically, which is incorrect.

These are known as the *Kramers-Kronig relations* on real and imaginary parts of χ.[19] If the real part of χ is known at all frequencies, we can use the Kramers-Kronig relations to generate the imaginary part, and vice versa. We see that the real and imaginary parts of χ cannot be chosen independently, if we are to respect the principle of causality.

Example 7.9

Show that the expression in square brackets of (7.80) is zero when $t' > t$, if $\chi(\omega)$ satisfies the Kramers-Kronig relations (7.85).

Solution: The expression may be written as

$$\begin{aligned}\int_{-\infty}^{\infty}\chi(\omega)\,e^{-i\omega(t-t')}\,d\omega &= \int_{-\infty}^{\infty}\mathrm{Re}\chi(\omega)\,e^{-i\omega(t-t')}\,d\omega + i\int_{-\infty}^{\infty}\mathrm{Im}\chi(\omega)\,e^{-i\omega(t-t')}\,d\omega\\ &= \int_{-\infty}^{\infty}\mathrm{Re}\chi(\omega)\,e^{-i\omega(t-t')}\,d\omega + i\int_{-\infty}^{\infty}\left[-\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\mathrm{Re}\chi(\omega')}{\omega'-\omega}\,d\omega'\right]e^{-i\omega(t-t')}\,d\omega\\ &= \int_{-\infty}^{\infty}\mathrm{Re}\chi(\omega)\,e^{-i\omega(t-t')}\,d\omega + \int_{-\infty}^{\infty}\mathrm{Re}\chi(\omega')\left[\frac{1}{i\pi}\int_{-\infty}^{\infty}\frac{e^{-i\omega(t-t')}}{\omega'-\omega}\,d\omega\right]d\omega'\end{aligned} \tag{7.86}$$

where we have invoked the Kramers-Kronig relation for $\mathrm{Im}\chi(\omega)$ (7.85) and interchanged the order of integration in the final expression. Since we are specifically considering future times $t' > t$, we have by (7.81)

$$\frac{1}{i\pi}\int_{-\infty}^{\infty}\frac{e^{-i\omega(t-t')}}{\omega'-\omega}\,d\omega = -e^{-i\omega'(t-t')}$$

Hence

$$\begin{aligned}\int_{-\infty}^{\infty}\chi(\omega)\,e^{-i\omega(t-t')}\,d\omega &= \int_{-\infty}^{\infty}\mathrm{Re}\chi(\omega)\,e^{-i\omega(t-t')}\,d\omega - \int_{-\infty}^{\infty}\mathrm{Re}\chi(\omega')\,e^{-i\omega'(t-t')}\,d\omega'\\ &= 0\end{aligned} \tag{7.87}$$

Finally, it is worth noting that the Kramers-Kronig relations also apply to the real and imaginary parts of the index of refraction (subtract one). [20]

---

[19] As with (7.81), the *principal value* of the integral must be calculated. If the integral is performed numerically, the sampling of points should straddle the singularity symmetrically. Separately, the integral on each side of $\omega' = \omega$ diverges, but with opposite sign.

$$n(\omega)-1=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\kappa(\omega')}{\omega'-\omega}d\omega' \quad \text{and} \quad \kappa(\omega)=-\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{n(\omega')-1}{\omega'-\omega}d\omega' \tag{7.88}$$

One can use the Kramers-Kronig relations to find the real part of the index from a measurement of absorption, if the measurement is done over a broad enough range of the spectrum. This is the most useful form of the Kramers-Kronig relations.

It is sometimes convenient to multiply the numerator and denominator inside the integrands of (7.88) by $\omega'+\omega$. Then noting that n is an even function and κ is an odd function allows us to dismiss either ω' or ω in the numerator and integrate[21] over positive frequencies only:

$$n(\omega)-1=\frac{2}{\pi}\int_{0}^{\infty}\frac{\omega'\kappa(\omega')}{\omega'^2-\omega^2}d\omega' \quad \text{and} \quad \kappa(\omega)=-\frac{2\omega}{\pi}\int_{0}^{\infty}\frac{n(\omega')-1}{\omega'^2-\omega^2}d\omega' \tag{7.89}$$

---

[20] This follows from Cauchy's theorem since the index (subtract one) is the square root of $\chi(\omega)$. The Kramers-Kronig relations for $\chi(\omega)$ guarantee that $\chi(\omega)$ has no poles in the upper half complex plane, when ω is considered (for mathematical purposes) to be a complex variable. Taking the square root does not introduce poles into the upper half plane.

[21] The integrals (7.88) and (7.89) diverge to either side of $\omega'=\omega$, but with opposite sign. Again, the principal value of the integral is required, which means a numeric grid should straddle the singularity symmetrically.

Exercises

Exercises for 7.1 Intensity of Superimposed Plane Waves

P7.1 (a) Consider two counter-propagating fields described by $\hat{\mathbf{x}}E_1 e^{i(kz-\omega t)}$ and $\hat{\mathbf{x}}E_2 e^{i(-kz-\omega t)}$ where E_1 and E_2 are both real. Show that their sum can be written as

$$\hat{\mathbf{x}}E_{\text{tot}}(z)\, e^{i(\Phi(z)-\omega t)}$$

where

$$E_{\text{tot}}(z) = E_1\sqrt{\left(1-\frac{E_2}{E_1}\right)^2 + 4\frac{E_2}{E_1}\cos^2 kz}$$

and

$$\Phi(z) = \tan^{-1}\left[\frac{(1-E_2/E_1)}{(1+E_2/E_1)}\tan kz\right]$$

Outside the range $-\frac{\pi}{2} \le kz \le \frac{\pi}{2}$ the pattern repeats.

(b) Suppose that two counter-propagating laser fields have separate intensities, I_1 and $I_2 = I_1/100$. The ratio of the fields is then $E_2/E_1 = 1/10$. In the standing interference pattern that results, what is the ratio of the maximum *intensity* to the minimum *intensity*? Are you surprised how high this is?

P7.2 Equation (7.7) implies that there is no *interference* between fields that are polarized along orthogonal dimensions. That is, the intensity of

$$\mathbf{E}(\mathbf{r},t) = \hat{\mathbf{x}}E_0 e^{i[(k\hat{\mathbf{z}})\cdot\mathbf{r}-\omega t]} + \hat{\mathbf{y}}E_0 e^{i[(k\hat{\mathbf{x}})\cdot\mathbf{r}-\omega t]}$$

according to (7.7) is uniform throughout space. Of course (7.7) does not apply since the **k**-vectors are not parallel. Show that the time-average of $\mathbf{S}(\mathbf{r},t)$ according to (7.4) exhibits interference in the distribution of net energy flow.

HINT: To get the B-field, see P1.2.

Exercises for 7.2 Group vs. Phase Velocity: Sum of Two Plane Waves

P7.3 Show that (7.10) can be written as

$$\mathbf{E}(\mathbf{r},t) = 2\mathbf{E}_0 e^{i\left(\frac{\mathbf{k}_2+\mathbf{k}_1}{2}\cdot\mathbf{r}-\frac{\omega_2+\omega_1}{2}t\right)}\cos\left(\frac{\Delta\mathbf{k}}{2}\cdot\mathbf{r}-\frac{\Delta\omega}{2}t\right)$$

From this show that the speed of the rapid-oscillation *intensity* peaks in Fig. 7.2 is $v_p = \bar{\omega}/\bar{k}$ where

$$\bar{\mathbf{k}} \equiv \frac{(\mathbf{k}_1+\mathbf{k}_2)}{2} \quad \text{and} \quad \bar{\omega} \equiv \frac{(\omega_1+\omega_2)}{2}$$

P7.4 Confirm the right-hand side of (7.17).

Exercises for 7.3 Frequency Spectrum of Light

P7.5 The continuous field of a very narrowband continuous laser may be approximated as a pure plane wave: $\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 e^{i(k_0 z - \omega_0 t)}$. Suppose the wave encounters a shutter at the plane $z = 0$.

(a) Compute the power spectrum of the light before the shutter. HINT: The answer is proportional to the square of a delta function centered on ω_0 (see (0.54)).

(b) Compute the power spectrum after the shutter if it is opened during the interval $-T/2 \leq t \leq T/2$. Plot the result. Are you surprised that the shutter appears to create extra frequency components?

HINT: Write your answer in terms of the sinc function defined by $\operatorname{sinc}\alpha \equiv \sin\alpha/\alpha$.

P7.6 (a) Consider the Gaussian pulse defined in (7.23). Determine the full width at half maximum (FWHM) of the intensity $I(\mathbf{r}, t)$, represented by T_{FWHM} (or Δt_{FWHM} if you wish), and FWHM of the power spectrum $I(\mathbf{r}, \omega)$, represented by Ω_{FWHM} (or $\Delta\omega_{\text{FWHM}}$ if you wish).

HINT: Both answers are in terms of T.

(b) Give an uncertainty principle for the product of $\Delta t_{\text{FWHM}} \Delta\omega_{\text{FWHM}}$.

Exercises for 7.5 Quadratic Dispersion

P7.7 The intensity of a laser pulse is Gaussian in time with a full width at half maximum duration $T_{\text{FWHM}} = 25$ fs and carrier frequency ω_0 corresponding to $\lambda_{\text{vac}} = 800$ nm. The pulse goes through a lens of thickness $\ell = 1$ cm (glass type BK7) with index of refraction given approximately by

$$n(\omega) \cong 1.4948 + 0.016\frac{\omega}{\omega_0}$$

What is the full width at half maximum duration of the intensity for the emerging pulse?

HINT: For the input pulse we have

$$T = \frac{T_{\text{FWHM}}}{2\sqrt{\ln 2}}$$

(see P7.6).

P7.8 If the pulse defined in (7.46) travels through the material for a very long distance z such that $\tilde{T}(z) \to T\Phi(z)$, show that the *instantaneous frequency* of the pulse (defined to be the time derivative of the overall phase) is

$$\omega_0 + \frac{t - z/v_g}{2\alpha z}$$

COMMENT: As the wave travels, the earlier part of the pulse oscillates more slowly than the later part. This is called chirp, and it means that the red frequencies get ahead of the blue ones since they experience a lower index. The instantaneous frequency is the effective local frequency.

Exercises for 7.6 Generalized Context for Group Delay

P7.9 When the spectrum of a pulse is narrow compared to the resonant spectral features of a material (like that depicted in Fig. 7.10), the reshaping delay (7.53) can be neglected. Show that the net delay in this case (7.52) reduces to

$$\lim_{T\to\infty} \Delta t_G(\mathbf{r}) = \left.\frac{\partial \mathrm{Re}\,\mathbf{k}}{\partial \omega}\cdot \Delta\mathbf{r}\right|_{\bar{\omega}}$$

HINT: The spectral intensity may be approximated as $I(\omega) = I_0\delta(\omega - \bar{\omega})$.

P7.10 When the spectrum a pulse is very broad, the reshaping delay (7.53) is negligible. Show that in this case the net delay reduces to

$$\lim_{T\to 0} \Delta t_G(\mathbf{r}) = \frac{\Delta r}{c}$$

assuming $\mathbf{k}$ and $\Delta\mathbf{r}$ are parallel. This implies that a sharply defined signal cannot travel faster than c.

HINT: The real index of refraction n goes to unity far from resonance, and the imaginary part κ goes to zero.

P7.11 Show that equation (7.49) can be written as

$$\langle t\rangle = -i\frac{\int\limits_{-\infty}^{\infty} d\omega \left[\frac{\partial}{\partial\omega}\mathbf{E}(\mathbf{r},\omega)\right]\cdot \mathbf{E}^*(\mathbf{r},\omega)}{\int\limits_{-\infty}^{\infty} d\omega \mathbf{E}(\mathbf{r},\omega)\cdot \mathbf{E}^*(\mathbf{r},\omega)} \equiv T\left[\mathbf{E}(\mathbf{r},\omega)\right]$$

HINT: Parseval's theorem (7.21) can be used in the denominator. In the numerator, substitute (7.18) for both fields, with ω and ω' as the dummy variables. Reorder integration to perform the time integral first. The following trick is handy:

$$\tfrac{1}{2\pi}\int\limits_{-\infty}^{\infty} t e^{-i(\omega'-\omega)t}dt = -i\tfrac{\partial}{\partial\omega}\tfrac{1}{2\pi}\int\limits_{-\infty}^{\infty} e^{-i(\omega'-\omega)t}dt = -i\tfrac{\partial}{\partial\omega}\delta\left(\omega'-\omega\right).$$

P7.12 Use the result of P 7.11 to derive (7.51).

HINT: Compute

$$\Delta t = T\left[\mathbf{E}(\mathbf{r}_0 + \Delta\mathbf{r},\omega)\right] - T\left[E(\mathbf{r}_0,\omega)\right].$$

and note that

$$\mathbf{E}\left(\mathbf{r}_0+\Delta\mathbf{r},\omega\right)=e^{i\mathbf{k}\cdot\Delta\mathbf{r}}\mathbf{E}\left(\mathbf{r}_0,\omega\right)=e^{i\mathrm{Re}[\mathbf{k}\cdot\Delta\mathbf{r}]}e^{-\mathrm{Im}[\mathbf{k}\cdot\Delta\mathbf{r}]}\mathbf{E}\left(\mathbf{r}_0,\omega\right)$$

The reshaping delay is

$$\Delta t_R\equiv T\left[e^{-\mathrm{Im}\mathbf{k}\cdot\Delta\mathbf{r}}E\left(\mathbf{r}_0,\omega\right)\right]-T\left[E\left(\mathbf{r}_0,\omega\right)\right]$$

The main effort is in showing

$$T\left[\mathbf{E}\left(\mathbf{r}_0+\Delta\mathbf{r},\omega\right)\right]=T\left[e^{-\mathrm{Im}[\mathbf{k}\cdot\Delta\mathbf{r}]}\mathbf{E}\left(\mathbf{r}_0,\omega\right)\right]+\frac{\int\limits_{-\infty}^{\infty}d\omega\left(\frac{\partial}{\partial\omega}\mathrm{Re}\mathbf{k}\cdot\Delta\mathbf{r}\right)\mathbf{E}\left(\mathbf{r}_0+\Delta\mathbf{r},\omega\right)\cdot\mathbf{E}^*\left(\mathbf{r}_0+\Delta\mathbf{r},\omega\right)}{\int\limits_{-\infty}^{\infty}d\omega\mathbf{E}\left(\mathbf{r}_0+\Delta\mathbf{r},\omega\right)\cdot\mathbf{E}^*\left(\mathbf{r}_0+\Delta\mathbf{r},\omega\right)}$$

Exercises for 7.A Pulse Chirping in a Grating Pair

P7.13 A Gaussian pulse with $T = 25$ fs is incident with $\theta_\mathrm{i} = 32°$ on a grating pair with groove separation $d = 0.833\ \mu$m. What grating separation L will lead to a pulse duration of $T = 100$ ps? Assume two passes through the grating pair for a total effective separation of $2L$. Take the pulse carrier frequency to corresponds to $\lambda_0 = 800$ nm.

Chapter 8

Coherence Theory

Coherence theory is the study of correlations that exist between different parts of a light field. *Temporal coherence* indicates a correlation between fields offset in time, $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{E}(\mathbf{r}, t-\tau)$. *Spatial coherence* has to do with correlations between fields at different spatial locations, $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{E}(\mathbf{r}+\Delta\mathbf{r}, t)$. Because light oscillations are too fast to resolve directly, we usually need to study optical coherence using interference techniques. In these techniques, light from different times or places in the light field are brought together at a detection point. If the two fields have a high degree of coherence, they consistently interfere either constructively or destructively at the detection point. If the two fields are not coherent, the interference at the detection point rapidly fluctuates between constructive and destructive interference, so that a time-averaged signal does not show interference.

You are probably already familiar with two instruments that measure coherence: the Michelson interferometer, which measures temporal coherence, and Young's two-slit interferometer, which measures spatial coherence. Your preliminary understanding of these instruments was probably gained in terms of single-frequency plane waves, which are perfectly coherent for all separations in time and space. In this chapter, we build on that foundation and derive descriptions that are appropriate when light with imperfect coherence is sent through these instruments. We also discuss a practical application known as Fourier spectroscopy (Section 8.4) which allows us to measure the spectrum of light using a Michelson interferometer rather than a grating spectrometer.

8.1 Michelson Interferometer

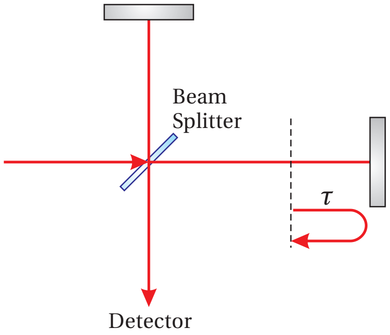


Figure 8.1 Michelson interferometer.

A Michelson interferometer employs a 50:50 beamsplitter to divide an initial beam into two identical beams and then delays one beam with respect to the other before bringing them back together (see Fig. 8.1). Depending on the relative path difference d (round trip by our convention) between the two arms of the system, the light can interfere constructively or destructively in the direction of the detector. The relative path difference d introduces a time delay τ, defined by

$\tau \equiv d/c$.

If the input light is a plane wave, the net field at the detector consists of the field coming from one arm of the interferometer $\mathbf{E}_0 e^{i(kz-\omega t)}$ added to the field coming from the other arm $\mathbf{E}_0 e^{i(kz-\omega(t-\tau))}$. These two fields are identical except for the delay τ. The intensity seen at the detector as a function of path difference is computed to be

$$\begin{aligned} I_{\text{tot}}(\tau) &= \frac{c\epsilon_0}{2}\left[\mathbf{E}_0 e^{i(kz-\omega t)} + \mathbf{E}_0 e^{i(kz-\omega(t-\tau))}\right]\cdot\left[\mathbf{E}_0 e^{i(kz-\omega t)} + \mathbf{E}_0 e^{i(kz-\omega(t-\tau))}\right]^* \\ &= \frac{c\epsilon_0}{2}\left[2\mathbf{E}_0\cdot\mathbf{E}_0^* + 2\mathbf{E}_0\cdot\mathbf{E}_0^*\cos(\omega\tau)\right] \\ &= 2I_0\left[1+\cos(\omega\tau)\right] \end{aligned}$$

(Plane Wave Input) (8.1)

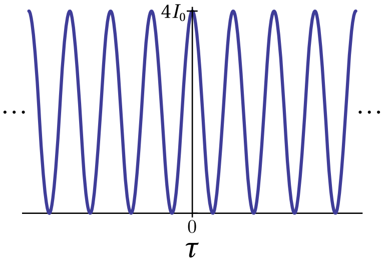


Figure 8.2 The intensity seen at the detector of a Michelson interferometer with a plane-wave input, as a function of the time delay τ. Because the plane wave is coherent over an infinite distance, the output oscillates without diminishing as the delay τ is adjusted in either direction. When the intensity at the detector is zero, all of the light is reflected back to the source.

where $I_0 \equiv \frac{c\epsilon_0}{2}\mathbf{E}_0\cdot\mathbf{E}_0^*$ is the intensity from one beam alone (when the other arm of the interferometer is blocked). This formula is probably familiar. It describes how the intensity at the detector oscillates between zero and four times the intensity from one beam,[1] as plotted in Fig. 8.2.

When light containing a continuous band of frequencies is sent through the interferometer, (8.1) no longer holds. Instead of repeating indefinitely, the oscillations at the detector become less pronounced as τ increases. The concept of *temporal coherence* describes how fast fringe visibility diminishes as delay is introduced in an arm of the Michelson interferometer. The less coherent the light source, the faster the fringes die out as the delay τ increases. To model this behavior, we need to expand our analysis beyond (8.1).

Consider an arbitrary waveform $\mathbf{E}(t)$ (comprised of many frequency components) that has traveled through the first arm of a Michelson interferometer to arrive at the detector in Fig. 8.1. The beam that travels through the second arm of the interferometer is identical, but delayed by the round-trip delay τ: $\mathbf{E}(t-\tau)$. The total field at the detector is the sum of these two fields:

$$\mathbf{E}_{\text{tot}}(t,\tau) = \mathbf{E}(t) + \mathbf{E}(t-\tau) \tag{8.2}$$

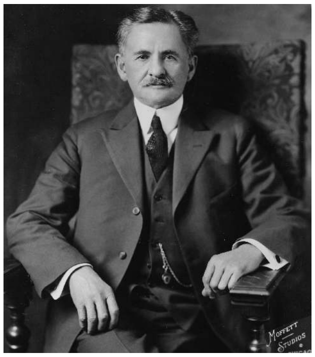
The total intensity I_{tot} at the detector is found using (7.20) (with $n = 1$):

$$\begin{aligned} I_{\text{tot}}(t,\tau) &= \frac{c\epsilon_0}{2}\mathbf{E}_{\text{tot}}(t,\tau)\cdot\mathbf{E}^*_{\text{tot}}(t,\tau) \\ &= \frac{c\epsilon_0}{2}\left[\mathbf{E}(t)\cdot\mathbf{E}^*(t) + \mathbf{E}(t)\cdot\mathbf{E}^*(t-\tau) + \mathbf{E}(t-\tau)\cdot\mathbf{E}^*(t) + \mathbf{E}(t-\tau)\cdot\mathbf{E}^*(t-\tau)\right] \\ &= I(t) + I(t-\tau) + \frac{c\epsilon_0}{2}\left[\mathbf{E}(t)\cdot\mathbf{E}^*(t-\tau) + \mathbf{E}(t-\tau)\cdot\mathbf{E}^*(t)\right] \\ &= I(t) + I(t-\tau) + c\epsilon_0\text{Re}\left\{\mathbf{E}(t)\cdot\mathbf{E}^*(t-\tau)\right\} \end{aligned} \tag{8.3}$$

As a reminder, the function $I(t) = \frac{c\epsilon_0}{2}\mathbf{E}(t)\cdot\mathbf{E}^*(t)$ corresponds to the intensity of the first beam at the detector when the second arm of the interferometer is blocked. The rapid oscillations of the light are automatically averaged away in $I(t)$ since we used (7.20), but the slowly varying envelope of the arbitrary pulse is retained. The intensity of the combined beams $I_{\text{tot}}(t,\tau)$ varies with t and also depends on the path delay τ.

---

[1] Keep in mind that if a 50:50 beam splitter is used, then the intensity arriving to the detector from one arm alone (with other arm blocked) is one fourth of the original beam, since the light meets the beam splitter twice.



Albert Abraham Michelson (1852–1931, United States) was born in Poland, but he immigrated to the US with his parents and grew up in the rough mining towns of California and Nevada where his father was a merchant. Michelson attended high school in San Francisco. He entered the US Naval Academy in 1869 (with intervention from US President Grant after Michelson pleaded his case when the president was walking near the White House). After two years at sea, Michelson returned to the Naval Academy to teach physics and mathematics for several years. Michelson was fascinated by the problem of determining the speed of light, and developed successive experiments to measure it more accurately. He is probably most famous for his experiment conducted at Case School of Applied Science in Cleveland with Edward Morley to detect the motion of the earth through the ether. Michelson later was a professor at the University of Chicago and then at Caltech. In 1907, he became the first American to win the Nobel prize, for his contributions to optics. Michelson married late in life and was the father of four. (Wikipedia)

We consider $\mathbf{I}(t)$ to be a pulse with a finite duration. We will be interested in the total amount of energy (per area) that the pulse deposits on a detector.[2] The detected signal, which we'll denote by $\text{Sig}(\tau)$, is the time-integrated intensity or *fluence*, having units of energy per area:

$$\text{Sig}(\tau) \propto \int_{-\infty}^{\infty} I_{\text{tot}}(t,\tau)\, dt \tag{8.4}$$

The proportionality accounts for the calibration of the detector, which might report in volts or current, etc. The fluence arriving at the detector is sensitive to the delay τ between the arms of the interferometer. Presumably, we can repeatedly send identical pulses into the interferometer and record $\text{Sig}(\tau)$ for many different delays τ. We can manipulate the fluence integral in (8.4) into a more useful form that will make the coherence properties more evident.

Manipulation of the fluence integral

Inserting (8.3) into the fluence integral, we have

$$\int_{-\infty}^{\infty} I_{\text{tot}}(t,\tau)\, dt = \int_{-\infty}^{\infty} I(t) dt + \int_{-\infty}^{\infty} I(t-\tau)\, dt + c\epsilon_0 \text{Re} \int_{-\infty}^{\infty} \mathbf{E}(t) \cdot \mathbf{E}^*(t-\tau)\, dt \tag{8.5}$$

The first two integrals on the right-hand side of (8.5) are equal,[3] and give the fluence $\mathscr{E}$ from either arm of the interferometer when the other arm is blocked:

$$\mathscr{E} \equiv \int_{-\infty}^{\infty} I(t) dt = \int_{-\infty}^{\infty} I(t-\tau)\, dt \tag{8.6}$$

The final integral in (8.5) remains unchanged if we take a Fourier transform followed by an inverse Fourier transform:

$$\int_{-\infty}^{\infty} \mathbf{E}(t) \cdot \mathbf{E}^*(t-\tau)\, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} d\omega e^{-i\omega\tau} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} d\tau e^{i\omega\tau} \int_{-\infty}^{\infty} \mathbf{E}(t) \cdot \mathbf{E}^*(t-\tau)\, dt \right] \tag{8.7}$$

The reason for this procedure is so that we can take advantage of the autocorrelation theorem described in P0.27. With it, the expression in square brackets simplifies to $\sqrt{2\pi}\mathbf{E}(\omega) \cdot \mathbf{E}^*(\omega) = \sqrt{2\pi} 2I(\omega)/c\epsilon_0$. Then with the aid of (8.6) and (8.7), the overall fluence (8.5) becomes

$$\int_{-\infty}^{\infty} I_{\text{tot}}(t,\tau)\, dt = 2\mathscr{E} \left[1 + \frac{1}{\mathscr{E}} \text{Re} \int_{-\infty}^{\infty} I(\omega) e^{-i\omega\tau} d\omega \right] \tag{8.8}$$

---

[2] For sub-nanosecond laser pulses, a detector automatically integrates the entire energy of the pulse since a detector cannot keep up with temporal variations on such a rapid time scale.

[3] Note that the second integral is insensitive to τ since a change of variables $t' = t - \tau$ converts it into the first integral.

With (8.8), we can rewrite the physical signal (8.4) in the more useful form

$$\mathrm{Sig}(\tau) \propto 2\mathscr{E}\left[1+\mathrm{Re}\left\{\gamma(\tau)\right\}\right] \tag{8.9}$$

where the dependence on the path delay τ is entirely contained in the *degree of coherence* function $\gamma(\tau)$:[4]

$$\gamma(\tau) \equiv \frac{\int\limits_{-\infty}^{\infty} I(\omega)\, e^{-i\omega\tau}\, d\omega}{\int\limits_{-\infty}^{\infty} I(\omega)\, d\omega} \tag{8.10}$$

The denominator of (8.10) was rewritten with the help of Parseval's theorem $\mathscr{E} \equiv \int_{-\infty}^{\infty} I(t)\,dt = \int_{-\infty}^{\infty} I(\omega)\,d\omega$. Remarkably, the signal out of the Michelson interferometer does not depend on the phase of $\mathbf{E}(\omega)$. It depends only on the amount of light associated with each frequency through $I(\omega) \equiv \frac{\epsilon_0 c}{2}\mathbf{E}(\omega)\cdot\mathbf{E}^*(\omega)$.

Alternate derivation of (8.9)

We could have derived (8.9) using another strategy, which may seem more intuitive than the approach above. Equation (8.1) gives the intensity at the detector when a single plane wave of frequency ω goes through the interferometer. Now suppose that a waveform composed of many frequencies is sent through the interferometer. The intensity associated with each frequency acts independently, obeying (8.1) individually.

The total energy (per area) accumulated at the detector is then a linear superposition of the spectral intensities of all frequencies present:

$$\int\limits_{-\infty}^{\infty} I_{\text{tot}}(\omega,\tau)\, d\omega = \int\limits_{-\infty}^{\infty} 2I(\omega)\left[1+\cos(\omega\tau)\right] d\omega \tag{8.11}$$

While this procedure may seem obvious, the fact that we can do it is remarkable! Remember that it is usually the fields that we must add together before finding the intensity of the resulting superposition. The formula (8.11) with its superposition of intensities relies on the fact that the different frequencies inside the interferometer when *time-averaged* (over all time) do not interfere. Certainly, the fields at different frequencies do interfere (or beat in time). However, they constructively interfere as often as they destructively interfere, and in a time-averaged picture it is as though the individual frequency components transmit independently. Again, in writing (8.11) we considered the light to be pulsed rather than continuous so that the integrals converge.

We can manipulate (8.11) as follows:

$$\int\limits_{-\infty}^{\infty} I_{\text{tot}}(\omega,\tau)\, d\omega = \left[2\int\limits_{-\infty}^{\infty} I(\omega)\, d\omega\right]\left[1+\frac{\int\limits_{-\infty}^{\infty} I(\omega)\cos(\omega\tau)\, d\omega}{\int\limits_{-\infty}^{\infty} I(\omega)\, d\omega}\right] \tag{8.12}$$

---

[4] M. Born and E. Wolf, *Principles of Optics*, 7th ed., p. 570 (Cambridge University Press, 1999).

This is the same as (8.8) since we can replace $\cos(\omega\tau)$ with $\mathrm{Re}\left\{e^{-i\omega\tau}\right\}$, and we can apply Parseval's theorem (8.6) to the other integrals. Thus, the above arguments lead to (8.9) and (8.10).

Example 8.1

Compute the output signal when a Gaussian pulse with spectrum (7.25) is sent into a Michelson interferometer.

Solution: The power spectrum of the pulse is[5]

$$I(\omega) = \frac{\epsilon_0 c}{2}\mathbf{E}_0 \cdot \mathbf{E}_0^* T^2 e^{-T^2(\omega-\omega_0)^2}$$

where T is the pulse duration, not to be confused with τ, the delay of the interferometer arm. As shown in Example 7.3, we also have

$$\int_{-\infty}^{\infty} I(\omega)\,d\omega = \frac{\epsilon_0 c}{2}\mathbf{E}_0 \cdot \mathbf{E}_0^* T\sqrt{\pi}$$

The degree of coherence (8.10) is then

$$\begin{aligned}\gamma(\tau) &= \frac{T}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-T^2(\omega-\omega_0)^2} e^{-i\omega\tau}\,d\omega \\ &= \frac{T}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-T^2\omega^2+(2T^2\omega_0-i\tau)\omega-T^2\omega_0^2}\,d\omega = \frac{T}{\sqrt{\pi}}\sqrt{\frac{\pi}{T^2}}\,e^{\frac{(2T^2\omega_0-i\tau)^2}{4T^2}-T^2\omega_0^2} \\ &= e^{-\frac{\tau^2}{4T^2}} e^{-i\omega_0\tau}\end{aligned}$$

Formula (0.55) was used to complete the integration. According to (8.9), the signal at the detector is then

$$\mathrm{Sig}(\tau) \propto 2\mathscr{E}\left[1+\mathrm{Re}\left\{\gamma(\tau)\right\}\right] = 2\mathscr{E}\left[1+e^{-\frac{\tau^2}{4T^2}}\cos(\omega_0\tau)\right]$$

Figure 8.3 shows this signal for a given T. As delay is added (or subtracted), the output signal oscillates. Eventually enough delay is introduced such that the very short pulses no longer interfere (arriving sequentially), and the output signal becomes steady.

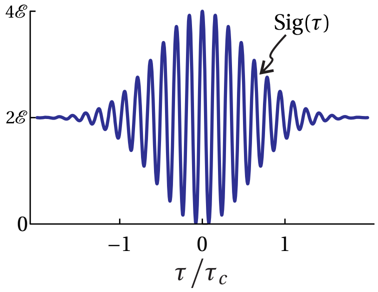


Figure 8.3 The output or signal from a Michelson interferometer for light with a Gaussian spectrum.

8.2 Coherence Time and Fringe Visibility

The degree of coherence function $\gamma(\tau)$ describes the oscillations in intensity at the detector as the mirror in one arm of the interferometer is moved. The real part of $\gamma(\tau)$ is analogous to $\cos(\omega\tau)$ in (8.1). However, for large delays τ, the oscillations tend to die off as different frequencies get out of sync—some interfere constructively, while others interfere destructively. Narrowband light is temporally more *coherent* than broadband light because there is less opportunity for frequencies to get out of sync. Still, for large path differences, the oscillations eventually die off, and the time-integrated intensity at the detector then remains steady as the mirror is moved further.

---

[5] Technically, the output intensity is one fourth this, but our calculation of the degree of coherence is insensitive to amplitude.

The *coherence time* τ_c is the amount of delay necessary to cause $\gamma(\tau)$ to quit oscillating (i.e. its amplitude approaches zero). This definition is not very precise, since the oscillations do not usually have an abrupt end, but instead slowly die off as τ increases. A useful (although arbitrary) analytic definition for the coherence time is

$$\tau_c \equiv \int_{-\infty}^{\infty} |\gamma(\tau)|^2 \, d\tau = 2\int_{0}^{\infty} |\gamma(\tau)|^2 \, d\tau \tag{8.13}$$

The *coherence length* is the distance that light travels in this time:

$$\ell_c \equiv c\tau_c \tag{8.14}$$

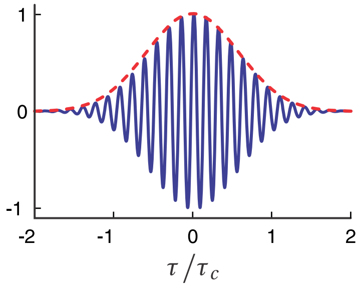


Figure 8.4 $\mathrm{Re}\{\gamma(\tau)\}$ (solid) and $|\gamma(\tau)|$ (dashed) for a light pulse with a Gaussian spectrum as in examples 8.1 and 8.2.

Another useful concept is *fringe visibility*. The fringe visibility is defined in the following way:

$$V(\tau) \equiv \frac{\max[\mathrm{Sig}(\tau)] - \min[\mathrm{Sig}(\tau)]}{\max[\mathrm{Sig}(\tau)] + \min[\mathrm{Sig}(\tau)]} \tag{8.15}$$

where $\max[\mathrm{Sig}(\tau)]$ refers to the detector signal when the mirror is positioned such that the amount of throughput to the detector is a local maximum, and $\min[\mathrm{Sig}(\tau)]$ refers to the detector signal when the mirror is positioned such that the amount of throughput to the detector is a local minimum. The minimum and the maximum don't occur at exactly the same τ, but the difference in τ is only about half an optical period. As the mirror moves a large distance from the equal-path-length position, the oscillations in $\mathrm{Sig}(\tau)$ become less pronounced as the max and min tend to the same value, and the fringe visibility goes to zero when $\gamma(\tau) = 0$. It is left as an exercise (see P8.1) to show that the fringe visibility can be written simply as[6]

$$V(\tau) = |\gamma(\tau)| \tag{8.16}$$

Note that the fringe visibility depends only upon the frequency content of the light without regard to whether the frequency components are organized into a short pulse or a longer time pattern.

Example 8.2

Find the fringe visibility and the coherence time for the Gaussian pulse studied in Example 8.1.

---

[6] M. Born and E. Wolf, *Principles of Optics*, 7th ed., p. 570 (Cambridge University Press, 1999).

Solution: By (8.16), the fringe visibility is

$$V(\tau) = |\gamma(\tau)| = e^{-\frac{\tau^2}{4T^2}}.$$

This is shown as the dashed line in Fig. 8.4. As expected, the fringe visibility dies off as delay τ gets farther from the origin (i.e. where the interferometer arms are equidistant). From (8.13) the coherence time is

$$\tau_c = \int_{-\infty}^{\infty} |\gamma(\tau)|^2 \, d\tau = \int_{-\infty}^{\infty} e^{-\frac{\tau^2}{2T^2}} \, d\tau = \sqrt{2\pi}\, T$$

which is the delay necessary to cause the fringes to substantially diminish.

8.3 Temporal Coherence of Continuous Sources

Consider a continuous light source such as starlight or a *continuous wave* (CW) laser. The integral $\int_{-\infty}^{\infty} I(t)\,dt$ diverges for such a source, since it is on forever (or at least for a very long time) and emits infinite (or very much) total energy. The concept of fluence (i.e. total energy) in this case is not very useful. However, note that the integrals on both sides of (8.5) diverge in the same way. We can renormalize (8.5) in this case by replacing the integrals on each side with the average value of the intensity:

$$I_{\text{ave}} \equiv \langle I(t)\rangle_t = \frac{1}{T}\int_{-T/2}^{T/2} I(t)\,dt \qquad \text{(continuous source)} \quad (8.17)$$

The duration T must be large enough to average over any fluctuations that are present in the light source.

For a continuous light source, the signal at the detector (8.9) becomes

$$\text{Sig}(\tau) \propto 2\langle I(t)\rangle_t \left[1 + \text{Re}\{\gamma(\tau)\}\right] \qquad \text{(continuous source)} \quad (8.18)$$

Although technically the integrals used in (8.10) to compute $\gamma(\tau)$ also diverge in the case of continuous light, the numerator and the denominator diverge in the same way. Therefore, we may renormalize $I(\omega)$ in a similar fashion to deal with this problem. The units in the numerator and denominator cancel so that $\gamma(\tau)$ always remains dimensionless. Once we have the degree of coherence function $\gamma(\tau)$, we can calculate the coherence time and fringe visibility just as we did for pulsed sources.

8.4 Fourier Spectroscopy

As we saw in (8.8), the signal output from a Michelson interferometer for a pulsed input may be written as

$$\text{Sig}(\tau) \propto 2\mathscr{E} + 2\text{Re}\int_{-\infty}^{\infty} I(\omega) e^{-i\omega\tau}\, d\omega \qquad (8.19)$$

Given a measurement of $\text{Sig}(\tau)$, we might like to find the power spectrum $I(\omega)$ that gave it. Unfortunately, $I(\omega)$ is buried within an integral in (8.19). However, since the integral looks like an inverse Fourier transform of $I(\omega)$, we will be able to extract the desired spectrum with a bit of work. This procedure for extracting $I(\omega)$ from an interferometric measurement is known as *Fourier spectroscopy*.[7]

Extracting $I(\omega)$

We first take the Fourier transform of (8.19):[8]

$$\mathscr{F}\left\{\text{Sig}(\tau)\right\} \propto \mathscr{F}\{2\mathscr{E}\} + \mathscr{F}\left\{2\text{Re}\int_{-\infty}^{\infty} I(\omega)\, e^{-i\omega\tau}\, d\omega\right\} \tag{8.20}$$

The left-hand side is known since it is the measured data, and a computer can be employed to take the Fourier transform of it. The first term on the right-hand side is the Fourier transform of a constant:

$$\mathscr{F}\{2\mathscr{E}\} = 2\mathscr{E}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{i\omega\tau}\, d\tau = 2\mathscr{E}\sqrt{2\pi}\delta(\omega) \tag{8.21}$$

Notice that (8.21) is zero everywhere except where $\omega = 0$, where a spike occurs. This represents the DC component of $\mathscr{F}\left\{\text{Sig}(\tau)\right\}$.

The second term of (8.20) can be written as

$$\mathscr{F}\left\{2\text{Re}\int_{-\infty}^{\infty} I(\omega)\, e^{-i\omega\tau}\, d\omega\right\} = \mathscr{F}\left\{\int_{-\infty}^{\infty} I(\omega)\, e^{-i\omega\tau}\, d\omega + \int_{-\infty}^{\infty} I(\omega)\, e^{i\omega\tau}\, d\omega\right\}$$

Carrying out the Fourier transforms gives

$$\int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} I(\omega')e^{-i\omega'\tau}\, d\omega'\right)e^{i\omega\tau}\, d\tau + \int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} I(\omega')e^{i\omega'\tau}\, d\omega'\right)e^{i\omega\tau}\, d\tau$$

which we rearrange to

$$\sqrt{2\pi}\left[\int_{-\infty}^{\infty} I(\omega')\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i(\omega'-\omega)\tau}\, d\tau\right)d\omega' + \int_{-\infty}^{\infty} I(\omega')\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i(\omega'+\omega)\tau}\, d\tau\right)d\omega'\right]$$

From (0.52) we note that the terms in parentheses are delta functions, so we have

$$\sqrt{2\pi}\left[\int_{-\infty}^{\infty} I(\omega')\delta\left(\omega'-\omega\right)d\omega' + \int_{-\infty}^{\infty} I(\omega')\delta\left(\omega'+\omega\right)d\omega'\right]$$

The remaining frequency integrals can then be easily performed to obtain

$$\mathscr{F}\left\{2\text{Re}\int_{-\infty}^{\infty} I(\omega)\, e^{-i\omega\tau}\, d\omega\right\} = \sqrt{2\pi}\left[I(\omega) + I(-\omega)\right] \tag{8.22}$$

---

[7] J. Peatross and S. Bergeson, "Fourier Spectroscopy of Ultrashort Laser Pulses," Am. J. Phys. **74**, 842-845 (2006).

[8] This is weird since normally we take Fourier transforms on fields rather than expressions involving intensity!

With (8.21) and (8.22) we can write (8.20) as

$$\mathscr{F}\left\{\operatorname{Sig}(\tau)\right\} \propto 2\mathscr{E}_0\delta(\omega) + I(\omega) + I(-\omega) \tag{8.23}$$

The Fourier transform of the measured signal is seen to contain three terms, one of which is the power spectrum $I(\omega)$ that we are after. Fortunately, when graphed as a function of ω (shown in Fig. 8.5), the three pieces on the right-hand side of (8.20) do not overlap. As a reminder, the measured signal as a function of τ looks something like that in Fig. 8.3. The oscillation frequency of the fringes lies in the neighborhood of ω_0. In summary, to obtain $I(\omega)$ using a Michelson interferometer, 1) record $\operatorname{Sig}(\tau)$; 2) take its Fourier transform; and 3) extract the curve at positive frequencies.

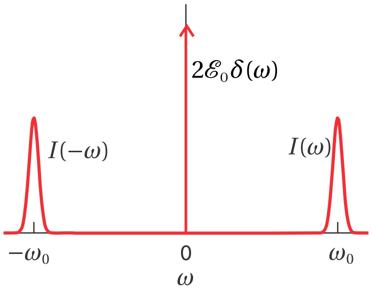


Figure 8.5 A graphical depiction of $\mathscr{F}\{\operatorname{Sig}(\tau)\}/\sqrt{2\pi}$.

8.5 Young's Two-Slit Setup and Spatial Coherence

In close analogy with the Michelson interferometer, which is useful for investigating temporal coherence, a *Young's two-slit setup* can be used to investigate *spatial coherence* of quasi-monochromatic light. Thomas Young, who lived nearly a century before Michelson, used his two-slit setup for the first conclusive demonstration that light propagates as a wave. The Young's double-slit setup and the Michelson interferometer have in common that two beams of light travel different paths and then interfere. In the Michelson interferometer, one path is delayed with respect to the other so that temporal effects can be studied. In the Young's two-slit setup, two laterally separate points of the same wave are compared as they are sent through two slits.

Depending on the coherence of the light entering each slit, the fringe pattern observed can exhibit good or poor visibility. Just as the Michelson interferometer is sensitive to the *spectral content* of light, the Young's two-slit setup is sensitive to the *spatial extent* of the light source illuminating the two slits. For example, if light from a distant star (restricted by a filter to a narrow spectral range) is used to illuminate a double-slit setup, the resulting interference pattern appearing on a subsequent screen shows good or poor fringe visibility depending on the angular width of the star. Michelson was the first to use this type of setup to measure the angular width of stars.

In contrast, light emerging from a single ideal point source has wavefronts that are spatially uniform in a lateral sense (see Fig. 8.6). Such wavefronts are said to be *spatially coherent*, even if the temporal coherence is not perfect (i.e. if a range of frequencies is present). When spatially coherent light illuminates a Young's two-slit setup, fringes of maximum visibility are seen at a distant screen, meaning the fringes vary between a maximum intensity and zero.

As a warmup exercise, we first consider a Young's two-slit setup illuminated by a single point source. Let the slits be equidistant from the point source. We represent the fields on a subsequent screen that transmit through each slit, respectively, as $\mathbf{E}_0 e^{i(kd_1-\omega t)}$ and $\mathbf{E}_0 e^{i(kd_2-\omega t)}$. The two fields are identical except for a phase associated with the distance from each slit to a particular point on the

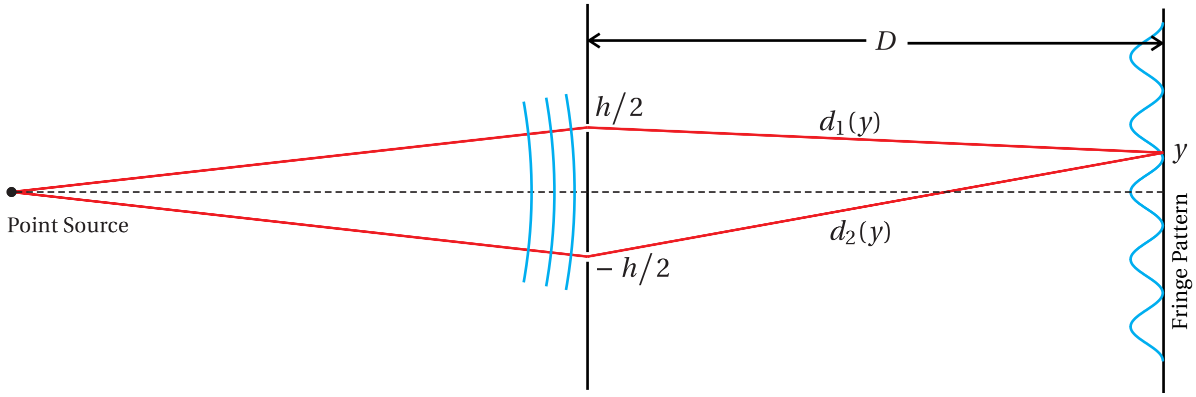


Figure 8.6 A point source produces coherent (locked phases) light. When this light which traverses two slits and arrives at a screen it produces a fringe pattern.

screen. In close analogy with (8.1), the resulting intensity pattern on a far-away screen is

$$I_{\text{tot}}(h) = 2I_0\left[1+\cos\left(kd_2 - kd_1\right)\right] = 2I_0\left[1+\cos\left(khy/D\right)\right] \tag{8.24}$$

Notice the close similarity between this expression and the output from a Michelson interferometer for a plane wave (8.1). We will consider h (the separation of the slits) to be the counterpart of τ (the delay introduced by moving a mirror in the Michelson interferometer). To obtain the final expression in (8.24) we made use of the following Taylor expansions:

$$d_1(y) = \sqrt{\left(y-h/2\right)^2 + D^2} = D\sqrt{1+\frac{\left(y-h/2\right)^2}{D^2}} \cong D\left[1+\frac{\left(y-h/2\right)^2}{2D^2}+\cdots\right] \tag{8.25}$$

and

$$d_2(y) = \sqrt{\left(y+h/2\right)^2 + D^2} = D\sqrt{1+\frac{\left(y+h/2\right)^2}{D^2}} \cong D\left[1+\frac{\left(y+h/2\right)^2}{2D^2}+\cdots\right] \tag{8.26}$$

These approximations are valid so long as $D \gg y$ and $D \gg h$.

We next consider how to modify (8.24) so that it applies to the case when the two slits are illuminated by a collection of point sources distributed over a finite lateral extent. This situation is depicted in Fig. 8.7 and it leads to *partial* spatial coherence if the phase of each point emitter fluctuates randomly. When a Young's two-slit setup is illuminated by an extended random source, the wavefronts at the two slits are less correlated. This makes the fringes move around on the screen rapidly and partially 'wash out' when time averaged, meaning worse fringe visibility.

To simplify our analysis, we restrict the distribution of point sources to vary only in the y' dimension.[9] We assume that the light is *quasi-monochromatic* so

---

[9] The results can be generalized to a two-dimensional source.

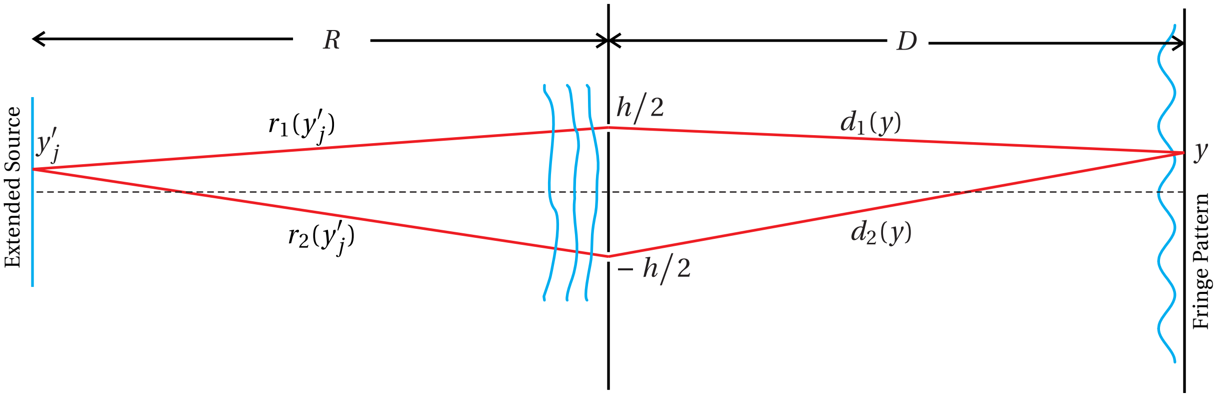


Figure 8.7 Light from an extended source is only partially coherent. Fringes are still possible, but they exhibit less contrast.

that its frequency is approximately ω with a phase that fluctuates randomly over time intervals much longer than the period of oscillation $2\pi/\omega$.[10]

The light emerging from the j^{th} point at y'_j travels by means of two very narrow slits to a point y on a screen. Let $\mathbf{E}_1(y'_j)$ and $\mathbf{E}_2(y'_j)$ be the fields on the screen at y, both originating from the point y'_j, but traveling respectively through the two different slits. We assume that these fields have the same polarization, and we will suppress the vectorial nature of the fields. For simplicity, we assume the two fields have the same (real) amplitude at the screen $E_0(y'_j)$. Thus, we write the two fields as

$$E_1(y'_j) = E_0(y'_j)e^{i\left\{k\left[r_1(y'_j)+d_1(y)\right]-\omega t+\phi(y'_j)\right\}} \tag{8.27}$$

and

$$E_2(y'_j) = E_0(y'_j)e^{i\left\{k\left[r_2(y'_j)+d_2(y)\right]-\omega t+\phi(y'_j)\right\}} \tag{8.28}$$

We have explicitly included an arbitrary phase $\phi(y'_j)$ assigned to each emission point at the source.

We now set about finding the cumulative field at y arising from the many points indexed by the subscript j. The total field on the screen at point y is

$$E_{\text{tot}}(h) = \sum_j \left[E_1(y'_j) + E_2(y'_j)\right] \tag{8.29}$$

Obviously, in addition to h, the total field depends on y, R, D, and k as well as on the phase $\phi(y'_j)$ at each point. Nevertheless, in the end we will mainly emphasize the dependence on the slit separation h. The intensity associated with (8.29) is

---

[10] Random phase fluctuations necessarily imply some frequency bandwidth, however small. Hence the need to specify *quasi-monochromatic light*.

$$\begin{aligned} I_{\text{tot}}(h) &= \frac{\epsilon_0 c}{2} \left| E_{\text{tot}}(h) \right|^2 \\ &= \frac{\epsilon_0 c}{2} \left[\sum_j E_1(y'_j) + E_2(y'_j) \right] \left[\sum_m E_1(y'_m) + E_2(y'_m) \right]^* \\ &= \frac{\epsilon_0 c}{2} \sum_{j,m} \left[E_1(y'_j) E_1^*(y'_m) + E_2(y'_j) E_2^*(y'_m) + E_1(y'_j) E_2^*(y'_m) + E_2(y'_j) E_1^*(y'_m) \right] \\ &= \frac{\epsilon_0 c}{2} \sum_{j,m} \left| E_0(y'_j) \right| \left| E_0(y'_m) \right| \left[e^{ik\left(r_1(y'_j) - r_1(y'_m)\right)} + e^{ik\left(r_2(y'_j) - r_2(y'_m)\right)} \right. \\ &\qquad \left. + 2\text{Re}\left\{ e^{ik\left(r_1(y'_j) - r_2(y'_m)\right)} e^{ik(d_1(y) - d_2(y))} \right\} \right] e^{i\left(\phi(y'_j) - \phi(y'_m)\right)} \end{aligned} \tag{8.30}$$

At this juncture we make a critical assumption: the phase of the emission $\phi(y'_j)$ varies in time independently at every point on the source. This is sometimes called the *stochastic* assumption, and it is appropriate for the emission from thermal sources such as starlight (filtered to a narrow frequency range), a glowing filament, or spontaneous emission from an excited gas or plasma. However, it is *not* appropriate for coherent sources like lasers (more on that in appendix 8.B).

A wonderful simplification happens to (8.30) when the phase difference $\phi(y'_j) - \phi(y'_m)$ varies randomly. If $j \neq m$, then $\exp\{i(\phi(y'_j) - \phi(y'_m))\}$ time-averages to zero. On the other hand, if $j = m$, then the factor reduces to $e^0 = 1$. Formally, this is written

$$\left\langle e^{i\left(\phi(y'_j) - \phi(y'_m)\right)} \right\rangle_t = \delta_{j,m} \equiv \begin{cases} 1 \text{ if } j = m, \\ 0 \text{ if } j \neq m. \end{cases} \qquad \text{(random phase assumption)} \tag{8.31}$$

where $\delta_{j,m}$ is known as the *Kronecker delta function*. The time-averaged intensity under the stochastic assumption (8.31) then reduces to

$$\langle I_{\text{tot}}(h) \rangle_t = \sum_j I(y'_j) + \sum_j I(y'_j) + 2\text{Re}\left\{ \sum_j I(y'_j) e^{ik\left(r_1(y'_j) - r_2(y'_j)\right)} e^{ik(d_1(y) - d_2(y))} \right\} \tag{8.32}$$

We may use (8.25) to simplify $d_1(y) - d_2(y) \cong hy/D$. Very similarly, we may also write $r_1(y'_j) - r_2(y'_j) \cong hy'_j/R$. The only thing left to do is to put (8.32) into a slightly more familiar form:

$$\langle I_{\text{tot}}(h) \rangle_t = \left[2 \sum_j I(y'_j) \right] \left[1 + \text{Re}\{\gamma(h)\} \right] \qquad \text{(random phase assumption)} \tag{8.33}$$

We have introduced

$$\gamma(h) \equiv \frac{e^{-i\frac{khy}{D}} \sum_j I(y'_j) e^{-i\frac{khy'_j}{R}}}{\sum_j I(y'_j)} \tag{8.34}$$

Thomas Young (1773–1829, English) was born in Milverton, Somerset, and was the oldest of ten children. By age fourteen, he had become proficient at a dozen different languages. As a young adult, he studied medicine and then went to Göttingen, Germany where he earned a doctoral degree in physics. In 1801, he was appointed professor of natural philosophy at the Royal Institute, but he also maintained an active medical practice on the side. He contributed to a wide variety of fields and helped to decipher ancient Egyptian hieroglyphs, including the Rosetta Stone. He published descriptions of the heart and arteries as well as how the eye accommodates to see at different depths and how the eye perceives color. In engineering fields, Young is well known his analysis of stresses and strains in elastic media. Young's double-slit experiment gave convincing evidence of the wave nature of light, overturning Newton's corpuscular theory. Regarding this, Thomas Young traded ideas with Augustin Fresnel through correspondence. (Wikipedia) which is known as the *degree of coherence*. It controls the fringe pattern seen at the screen.

We can generalize (8.33) so that it applies to the case of a continuous distribution of light as opposed to a collection of discrete point sources. In Appendix 8.A we show how summations in (8.33) and (8.34) become integrals over the source intensity distribution, and we write

$$\langle I_{\text{net}}(h)\rangle_t = 2\langle I_{\text{oneslit}}\rangle_t\left[1+\text{Re}\left\{\gamma(h)\right\}\right] \qquad \text{(random phase assumption)} \tag{8.35}$$

where

$$\gamma(h) \equiv \frac{e^{-i\frac{khy}{D}}\int\limits_{-\infty}^{\infty} I(y')e^{-i\frac{khy'}{R}}dy'}{\int\limits_{-\infty}^{\infty} I(y')dy'} \tag{8.36}$$

Here $I(y')$ has units of intensity per length of source.

The factor $\exp(-ikhy/D)$ defines the locations of the periodic fringes on the screen. The rest of (8.36) controls the (more interesting) depth of the fringes as the slit separation h is varied. When the slit separation h increases, the amplitude of $\gamma(h)$ tends to diminish until the intensity at the screen becomes uniform. When the two slits have very small separation (such that $e^{-i\frac{khy'}{R}} \cong 1$) then we have $|\gamma(h)| = 1$ and very good fringe visibility results. $\gamma(h)$ dictates the degree of *spatial coherence* in much the same way that $\gamma(\tau)$ dictates the degree of *temporal coherence*. Notice the close similarity between (8.36) and (8.10).

As the slit separation h increases, the fringe visibility

$$V(h) = |\gamma(h)| \tag{8.37}$$

diminishes, eventually approaching zero (see (8.16)). In analogy to the temporal case (see (8.13)), we can define a slit separation sufficiently large to make the fringes at the screen 'wash out':

$$h_c \equiv 2\int\limits_0^{\infty} |\gamma(h)|^2\, dh \tag{8.38}$$

Appendix 8.A Spatial Coherence for a Continuous Spatial Distribution

In this appendix we examine the spatial coherence of light from a *continuous* spatial distribution (as opposed to a collection of discrete point sources) and justify (8.36) and (8.37). We begin by replacing the summations in (8.30) with integrals over a continuous emission source. We make the following replacements:

$$\begin{aligned}
\sum_j E_1(y'_j) &\to \int_{-\infty}^{\infty} E_1(y')dy' \quad \text{and} \quad \sum_m E_1(y'_m) \to \int_{-\infty}^{\infty} E_1(y'')dy'' \\
\sum_j E_2(y'_j) &\to \int_{-\infty}^{\infty} E_2(y')dy' \quad \text{and} \quad \sum_m E_2(y'_m) \to \int_{-\infty}^{\infty} E_2(y'')dy''
\end{aligned} \tag{8.39}$$

Rather than deal with a time average of randomly varying phases, we will instead work with a linear superposition of all conceivable phase factors. That is, we will write the phase $\phi(y')$ as Ky', where K is a parameter with units of inverse length, which we allow to take on all possible real values with uniform likelihood. The way we modify (8.31) for the continuous case is then

$$\left\langle e^{i\left[\phi(y'_j)-\phi(y'_m)\right]}\right\rangle_t = \delta_{j,m} \to \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{iK(y'-y'')}dK = \delta(y''-y') \tag{8.40}$$

With the replacements in (8.39) and (8.40), (8.30) becomes

$$\begin{aligned}
I_{\text{tot}}(h) = \frac{\epsilon_0 c}{2}\int_{-\infty}^{\infty} dy'\left|E(y')\right|\int_{-\infty}^{\infty} dy''\left|E(y'')\right|\Big[&e^{ik\left(r_1(y')-r_1(y'')\right)} + e^{ik\left(r_2(y')-r_2(y'')\right)} \\
&+2\text{Re}\left\{e^{ik\left(d_1(y)-d_2(y)\right)}e^{ik\left(r_1(y')-r_2(y'')\right)}\right\}\Big]\delta(y''-y')
\end{aligned} \tag{8.41}$$

Again, consistent with (8.25), we may write $d_1(y) - d_2(y) \cong hy/D$ and $r_1(y') - r_2(y') \cong hy'/R$, and (8.41) reduces to

$$I_{\text{tot}}(h) = 2\int_{-\infty}^{\infty} I(y')dy' + 2\text{Re}\left\{e^{-i\frac{khy}{D}}\int_{-\infty}^{\infty} I(y')e^{-i\frac{khy'}{R}}dy'\right\} \tag{8.42}$$

where

$$I(y') \equiv \frac{1}{2}\epsilon_0 c\left|E(y')\right|^2 \tag{8.43}$$

For I_{tot} to have normal units of intensity, $I(y')$ must have units of intensity per length of source, implying that $E(y')$ has units of field per square root of length. Hence, $\int_{-\infty}^{\infty} I(y')dy'$ is the intensity at the screen caused by the entire extended source when only one slit is open. We see that (8.42) is equivalent to (8.35) and (8.36).

Appendix 8.B Van Cittert-Zernike Theorem

In this appendix we avoid making the assumption of randomly varying phase. This would be the case when the source of light is, for example, a laser. In place of (8.41) we have

$$\begin{aligned} I_{\text{tot}}(h) \propto & \left| \int_{-\infty}^{\infty} \left[\left| E(y') \right| e^{i\phi(y') + i\frac{ky'^2}{2R}} \right] e^{-i\frac{khy'}{2R}} dy' \right|^2 + \left| \int_{-\infty}^{\infty} \left[\left| E(y') \right| e^{i\phi(y') + i\frac{ky'^2}{2R}} \right] e^{i\frac{khy'}{2R}} dy' \right|^2 \\ & + 2\text{Re} e^{i\frac{khy}{D}} \left\{ \int_{-\infty}^{\infty} \left[\left| E(y') \right| e^{i\phi(y') + i\frac{ky'^2}{2R}} \right] e^{-i\frac{khy'}{2R}} dy' \right\} \left\{ \int_{-\infty}^{\infty} \left[\left| E(y') \right| e^{i\phi(y') + i\frac{ky'^2}{2R}} \right] e^{i\frac{khy'}{2R}} dy' \right\}^* \end{aligned} \tag{8.44}$$

where we have employed (8.25) and (8.26) and similar expressions involving R and y'.

The first term on the right-hand side of (8.44) is the intensity on the screen when the lower slit is covered. The second term is the intensity on the screen when the upper slit is covered. The last term is the interference term, which modifies the sum of the individual intensities when light goes through both slits.

Notice the occurrence of Fourier transforms (over position) on the quantities inside of the square brackets. Later, when we study diffraction theory, we will recognize these transforms as determining the strength of fields impinging on the individual slits. This corresponds to a major difference between a spatially coherent source and a random-phase source. With the random-phase source, the slits are always illuminated with the same strength regardless of the separation. However, with a coherent source, 'beaming' can occur such that the *strength* as well as phase of the field at each slit depends on the slit separation.

A beautiful simplification occurs when the phase of the emitted light has the following distribution:

$$\phi(y') = -\frac{ky'^2}{2R} \qquad \text{(converging spherical wave)} \tag{8.45}$$

Equation (8.45) is not as arbitrary as it may first appear. This particular phase is an approximation to a concave spherical wavefront converging to the center between the two slits. This type of wavefront is created when a plane wave passes through a lens. With the special phase (8.45), the intensity (8.44) reduces to

$$I_{\text{tot}}(h) \propto 2 \left| \int_{-\infty}^{\infty} \left| E(y') \right| e^{-i\frac{khy'}{2R}} dy' \right|^2 \left[1 + \text{Re}\left\{ e^{i\frac{khy}{D}} \right\} \right] \qquad \text{(converging spherical wave)} \tag{8.46}$$

The factor

$$\left| \int_{-\infty}^{\infty} \left| E(y') \right| e^{-i\frac{khy'}{2R}} dy' \right|$$

corresponds to the field impinging on the screen and which arises from either slit, when positioned at $h/2$. Let this field be denoted by $|E_1(h/2)|$. The field strength when the single slit is positioned at h compared to that when it is positioned at zero is

$$\left|\frac{E_1(h)}{E_1(0)}\right| = \left|\frac{\int\limits_{-\infty}^{\infty} \left|E(y')\right| e^{-i\frac{khy'}{R}} dy'}{\int\limits_{-\infty}^{\infty} \left|E(y')\right| dy'}\right| \quad \text{(converging spherical wave assumption)} \tag{8.47}$$

This looks very much like fringe visibility $\left|\gamma(h)\right|$ given by (8.37) and (8.36) except that the magnitude of the field appears in (8.47), whereas the intensity appears in (8.36).

This may seem rather contrived, but at least it is cute, and it is known as the van Cittert-Zernike theorem.[11] It says that the spatial coherence of an extended source with randomly varying phase drops off with lateral slit separation in the same way that the field pattern at the focus of a converging spherical wave would drop off, whose *field amplitude* distribution is the same as the original intensity distribution.

---

[11] M. Born and E. Wolf, *Principles of Optics*, 7th ed., p. 574 (Cambridge University Press, 1999).

Exercises

Exercises for 8.2 Coherence Time and Fringe Visibility

P8.1 (a) Verify that (8.16) gives the fringe visibility.

HINT: Write $\gamma = \left|\gamma\right| e^{i\phi}$ and assume that $\left|\gamma\right|$ varies slowly in comparison to the oscillations.

(b) What is the coherence time τ_c of the light in P8.4?

P8.2 (a) Show that the fringe visibility of a Gaussian spectral distribution (see Example 8.2) goes from 1 to $e^{-\pi/2} = 0.21$ as the round-trip delay increases from zero to the coherence length.

(b) Derive an expression for the FWHM wavelength bandwidth $\Delta\lambda_{\text{FWHM}}$ in terms of the coherence length ℓ_c and the center wavelength λ_0.

HINT: First determine $\Delta\omega_{\text{FWHM}}$, defined to be the width of $I(\omega)$ at half of its peak (see P7.6). To convert to a wavelength difference, use $\omega = \frac{2\pi c}{\lambda} \Rightarrow |\Delta\omega_{\text{FWHM}}| \cong \frac{2\pi c}{{\lambda_0}^2}\Delta\lambda_{\text{FWHM}}$.

Exercises for 8.3 Temporal Coherence of Continuous Sources

P8.3 Show that $\text{Re}\{\gamma(\tau)\}$ defined in (8.10) reduces to $\cos(\omega_0\tau)$ in the case of a plane wave $E(t) = E_0 e^{i(k_0 z - \omega_0 t)}$ being sent through a Michelson interferometer. In other words, the output intensity from the interferometer reduces to

$$I = 2I_0\left[1 + \cos(\omega_0\tau)\right]$$

as you already expect.

HINT: Don't be afraid of delta functions. After integration, the left-over delta functions cancel.

P8.4 Light emerging from a dense hot gas has a collisionally broadened power spectrum described by the Lorentzian function

$$I(\omega) = \frac{I(\omega_0)}{1 + \left(\frac{\omega - \omega_0}{\Delta\omega_{\text{FWHM}}/2}\right)^2}$$

The light is sent into a Michelson interferometer. Make a graph of the average intensity arriving to the detector as a function of τ.

HINT: See (0.56). You do not need to worry about the time average in (8.17); $I(\omega_0)$ can be thought of as already being normalized to the average intensity.

P8.5 (a) The spectral phase of the light in P8.4 is randomly organized. Describe qualitatively how the light behaves as a function of time.

(b) Now suppose that the phase of the light is somehow neatly organized such that

$$E(\omega) = \frac{iE(\omega_0)\, e^{i\frac{\omega}{c}z}}{i + \frac{\omega - \omega_0}{\Delta\omega_{\mathrm{FWHM}}/2}}$$

Perform the inverse Fourier transform on the field and determine the light *intensity* as a function of time. Make a sketch.

HINT:

$$\int_{-\infty}^{\infty} \frac{e^{-iax}}{x+\beta}\, dx = \begin{cases} -2i\pi e^{ia\beta} & \text{if } a>0 \\ 0 & \text{if } a<0 \end{cases} \qquad (\mathrm{Im}\beta > 0)$$

The constants $I(\omega_0)$, and $\Delta\omega_{\mathrm{FWHM}}$ will appear in the answer.

(c) Will the fringe visibility as a function of τ for the pulse in part (b) behave differently from the visibility for the continuous light in part (a)? Explain.

Exercises for 8.4 Fourier Spectroscopy

L8.6 (a) Use a scanning Michelson interferometer to measure the wavelength of the ultrashort laser pulses from a mode-locked Ti:sapphire oscillator.[12]

(b) Measure the coherence length of the source by observing the distance over which the visibility diminishes. Determine the bandwidth $\Delta\lambda_{\mathrm{FWHM}}$ of the source, assuming the Gaussian profile in P8.2.

(c) Use a computer to perform a fast Fourier transform (FFT) of the signal output. For the positive frequencies, plot the laser spectrum as a function of λ and compare with the results of (a) and (b).

(d) How do the results change if the ultrashort pulses are first stretched in time by traversing a thick piece of glass?

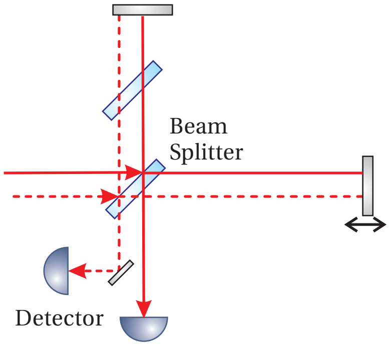


Figure 8.8

Exercises for 8.5 Young's Two-Slit Setup and Spatial Coherence

P8.7 (a) A point source with wavelength $\lambda = 500$ nm illuminates two parallel slits separated by $h = 1.0$ mm. If the screen is $D = 2$ m away, what is the separation between the interference peaks on the screen? Make a sketch.

---

[12] J. Peatross and S. Bergeson, "Fourier Spectroscopy of Ultrashort Laser Pulses," Am. J. Phys. **74**, 842-845 (2006).

(b) A thin piece of glass with thickness $d = 0.01$ mm and index $n = 1.5$ is placed in front of one of the slits. By how many fringes does the pattern at the screen move?

HINT: Add $\Delta\phi$ to $k(d_2 - d_1)$ in (8.24), where $\Delta\phi$ is the phase difference between traversing the glass and traversing an empty region of the same thickness.

L8.8 (a) Carefully measure the separation of a double slit in the lab ($h \sim 0.1$ mm separation) by shining a HeNe laser ($\lambda = 633$ nm) through it and measuring the interference peak separations on a distant wall (say, 2 m from the slits).

HINT: For better accuracy, measure across several fringes and divide.

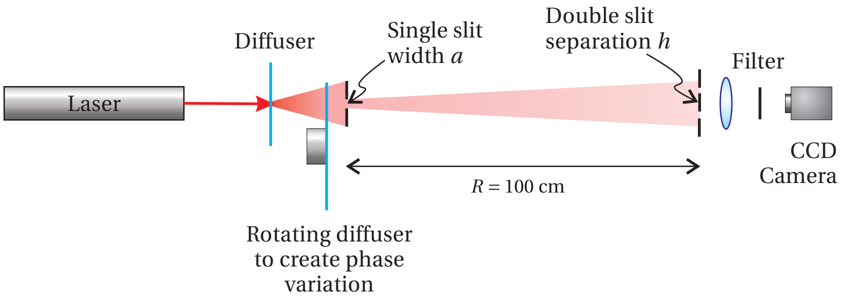

Laser
Diffuser
Single slit width a
Double slit separation h
Filter
CCD Camera
$R = 100$ cm
Rotating diffuser to create phase variation

Figure 8.9

(b) Create an extended light source with a HeNe laser using a time-varying diffuser followed by an adjustable single slit. (The diffuser must rotate rapidly to create random time variation of the phase at each point as would occur automatically for a natural source such as a star.) Place the double slit at a distance of $R \approx 100$ cm after the first slit. (Take note of the exact value of R, as you will need it for the next problem.) Use a lens to image the interference pattern that would have appeared on a far-away screen into a video camera. Observe the visibility of the fringes. Adjust the width of the source with the single slit until the visibility of the fringes disappears. After making the source wide enough to cause the fringe pattern to degrade, measure the single slit width a by shining a HeNe laser through it and observing the interference pattern on the distant wall. (video)

HINT: As we will study later, a single slit of width a produces an intensity pattern on a screen a distance L away described by

$$I(x) = I_{\text{peak}}\text{sinc}^2\left(\frac{\pi a}{\lambda L}x\right)$$

where $\text{sinc}(\alpha) \equiv \frac{\sin\alpha}{\alpha}$ and $\lim\limits_{\alpha\to 0} \frac{\sin\alpha}{\alpha} = 1$.

NOTE: It would have been nicer to vary the separation of the two slits to determine the width of a fixed source. However, because it is hard to make an adjustable double slit, we vary the size of the source until the spatial coherence of the light matches the slit separation.

P8.9 (a) Compute h_c for a uniform intensity distribution of width a using (8.38).

(b) Use this formula to check that your measurements in L8.8 agree with spatial coherence theory.

HINT: In your experiment h_c is the double slit separation. Use your measured R and h to calculate what the width of the single slit (i.e. a) should have been when the fringes disappeared and compare this calculation to your direct measurement of a.

Solution: (This is only a partial solution)

$$\gamma(h) = \frac{\int\limits_{-a/2}^{a/2} I_0 \exp\left[-ikh\left(\frac{y'}{R}+\frac{y}{D}\right)\right] dy'}{\int\limits_{-a/2}^{a/2} I_0 dy'} = \frac{e^{-ikh\frac{y}{D}} \int\limits_{-a/2}^{a/2} e^{-ikh\frac{y'}{R}} dy'}{a} = \frac{e^{-ikh\frac{y}{D}} \left[\frac{e^{-ikh\frac{y'}{R}}}{-i\frac{kh}{R}}\right]_{-a/2}^{a/2}}{a}$$

$$= e^{-ikh\frac{y}{D}} \left[\frac{e^{-ikh\frac{a/2}{R}} - e^{-ikh\frac{-a/2}{R}}}{-2ikh\frac{a/2}{R}}\right] = e^{-ikh\frac{y}{D}} \operatorname{sinc}\frac{kha}{2R}$$

Note that

$$\int\limits_0^\infty \frac{\sin^2 \alpha x}{(\alpha x)^2} dx = \frac{\pi}{2\alpha}$$

Review, Chapters 5–8

True and False Questions

R26 T or F: As light enters a crystal, the Poynting vector always obeys Snell's law.

R27 T or F: As light enters a crystal, the **k**-vector obeys Snell's law for the extraordinary wave.

R28 T or F: In our notation (widely used), $I(t)$ is the Fourier transform of $I(\omega)$.

R29 T or F: The integral of $I(t)$ over all t equals the integral of $I(\omega)$ over all ω.

R30 T or F: The phase velocity of light (the speed of an individual frequency component of the field) never exceeds the speed of light c.

R31 T or F: The group velocity of light can exceed c if absorption or amplification takes place.

R32 T or F: A Michelson interferometer is ideal for measuring the *spatial* coherence of light.

R33 T or F: A Michelson interferometer can be used to measure the spectral intensity of light $I(\omega)$.

R34 T or F: A Michelson interferometer can be used to measure the duration of a short laser pulse and thereby characterize its chirp.

R35 T or F: A Michelson interferometer can be used to measure the wavelength of light.

R36 T or F: A Michelson interferometer can be used to measure the phase of $E(\omega)$.

R37 T or F: The Fourier transform (or inverse Fourier transform if you prefer) of $I(\omega)$ is proportional to the degree of temporal coherence.

R38 T or F: The Young's two-slit setup is ideal for measuring the *temporal* coherence of light.

R39 T or F: Vertically polarized light illuminates a Young's double-slit setup and fringes are seen on a distant screen with good visibility. A half-wave plate is placed in front of one of the slits so that the polarization for that slit becomes horizontally polarized. **Statement:** The fringes at the screen will shift position but maintain their good visibility.

Problems

R40 Second harmonic generation (the conversion of light with frequency ω into light with frequency 2ω) can occur when very intense laser light travels in a material. For energy conversion, the laser light and the second harmonic light need to travel at the *same* speed in the material. In other words, both frequencies need to have the same index of refraction.

Unfortunately, the index of refraction is almost never the same for different frequencies in a given material, owing to dispersion. However, in some crystals when one frequency propagates as an ordinary wave and the other propagates as an extraordinary wave, the two indices can be made precisely the same by 'tuning' the angle of the crystal.

Consider a uniaxial KDP crystal (potassium dihydrogen phosphate) with ordinary index n_o and extraordinary index

$$\frac{n_o n_e}{\sqrt{n_o^2 \sin^2\theta + n_e^2 \cos^2\theta}}$$

where θ is the angle made with the optic axis. At the frequency of a ruby laser, KDP has indices $n_o(\omega) = 1.505$ and $n_e(\omega) = 1.465$. At the frequency of the second harmonic, the indices are $n_o(2\omega) = 1.534$ and $n_e(2\omega) = 1.487$.

In order to make the indices at the two frequencies the same, decide which frequency should propagate as an ordinary wave and which should propagate as an extraordinary one. What angle θ will make the indices the same?

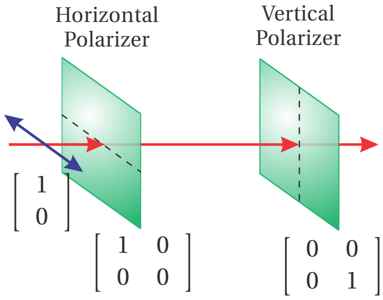

Horizontal Polarizer
Vertical Polarizer
$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$
$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$

Figure 8.10

R41 (a) Horizontally polarized light travels through a horizontal polarizer and then a vertical polarizer as shown. What is the Jones vector of the transmitted field?

(b) Now a polarizer at 45° is inserted between the two polarizers in the system described in part (a). What is the Jones vector of the transmitted field? How does the final intensity compare to initial intensity?

NOTE: The polarizer matrix is $\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.

(c) Now a quarter-wave plate with a fast-axis angle at 45° is inserted between the two polarizers (instead of the polarizer in part (b)). What is the Jones vector of the transmitted field? How does the final intensity compare to initial intensity? NOTE: The wave-plate matrix is $\frac{1}{2}\begin{bmatrix} 1+i & 1-i \\ 1-i & 1+i \end{bmatrix}$.

R42 (a) Derive the Jones matrix for a half-wave plate with its fast axis making an arbitrary angle θ with the x-axis.

HINT: Project an arbitrary polarization with E_x and E_y onto the fast and slow axes of the wave plate. Shift the slow axis phase by π, and then project the field components back onto the horizontal and vertical axes. The answer is

$$\begin{bmatrix} \cos^2\theta - \sin^2\theta & 2\sin\theta\cos\theta \\ 2\sin\theta\cos\theta & \sin^2\theta - \cos^2\theta \end{bmatrix}$$

(b) We desire to create a variable attenuator for a polarized laser beam using a half-wave plate and a polarizer aligned to the initial horizontal polarization of a beam (see Fig. 8.11). What is the ratio of the intensity exiting the polarizer to the incoming intensity as a function of θ?

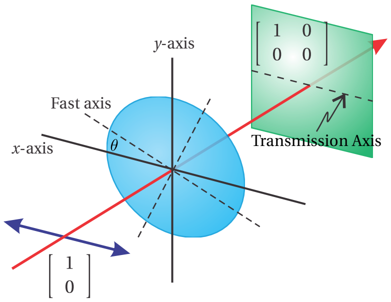

$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
y-axis
Fast axis
x-axis
θ
Transmission Axis
$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$

Figure 8.11 Polarizing Elements

R43 (a) What is the spectral content (i.e. $I(\omega)$) of a square laser pulse

$$E(t) = \begin{cases} E_0 e^{-i\omega_0 t} & , \quad |t| \le \tau/2 \\ 0 & , \quad |t| > \tau/2 \end{cases}$$

Make a sketch of $I(\omega)$, indicating the location of the first zeros.

(b) What is the temporal shape (i.e. $I(t)$) of a light pulse with frequency content

$$E(\omega) = \begin{cases} E_0 & , \quad |\omega - \omega_0| \le \Delta\omega/2 \\ 0 & , \quad |\omega - \omega_0| > \Delta\omega/2 \end{cases}$$

where in this case E_0 has units of E-field per frequency. Make a sketch of $I(t)$, indicating the location of the first zeros.

(c) If $E(\omega)$ is given (not necessarily the same function as above), and the light passes through a material with index $n(\omega)$ and thickness ℓ, how would you find $E(t)$ after exiting the material? Please set up the integral without performing it.

R44 (a) Prove Parseval's theorem:

$\int\limits_{-\infty}^{\infty} |E(\omega)|^2\, d\omega = \int\limits_{-\infty}^{\infty} |E(t)|^2\, dt.$ HINT: $\delta\left(t' - t\right) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} e^{i\omega(t'-t)}\, d\omega$

(b) Suppose that you have a detector that measures the total energy in a pulse of light, say 1 mJ directed onto an area of 0.01 cm^2. Next you measure the spectrum of light and find it to have a width of $\Delta\lambda =$

50 nm, centered at $\lambda_0 = 800$ nm. Assume that the light has a Gaussian frequency profile

$$I(\omega) = I(\omega_0)\, e^{-\left(\frac{\omega-\omega_0}{\Delta\omega}\right)^2}$$

Find the value and correct units for $I(\omega_0)$.

HINT: Use as an approximate value $\Delta\omega \cong \frac{2\pi c}{\lambda_0^2}|\Delta\lambda|$. Also

$$\int_{-\infty}^{\infty} e^{-Ax^2+Bx+C}\,dx = \sqrt{\frac{\pi}{A}}\, e^{B^2/4A+C} \qquad \mathrm{Re}\{A\} > 0$$

R45 Continuous light entering a Michelson interferometer has a spectrum described by

$$I(\omega) = \begin{cases} I_0 & , \quad |\omega-\omega_0| \le \Delta\omega/2 \\ 0 & , \quad |\omega-\omega_0| > \Delta\omega/2 \end{cases}$$

The Michelson interferometer uses a 50:50 beam splitter. The emerging light produces a signal $\mathrm{Sig}(t,\tau) \propto 1 + \mathrm{Re}\gamma(\tau)$, where degree of coherence is

$$\gamma(\tau) = \int_{-\infty}^{\infty} I(\omega)\, e^{-i\omega\tau}\, d\omega \Bigg/ \int_{-\infty}^{\infty} I(\omega)\, d\omega$$

Find the fringe visibility $V \equiv (I_{\max} - I_{\min})/(I_{\max} + I_{\min})$ as a function of τ (i.e. the round-trip delay due to moving one of the mirrors).

R46 A chirped Gaussian pulse has the form

$$E(t) = \frac{E_0}{\left(1+\Phi^2\right)^{\frac{1}{4}}}\, e^{-\frac{t^2(1+i\Phi)}{2T^2\left(1+\Phi^2\right)}}\, e^{-i\omega_0 t + i\frac{1}{2}\tan^{-1}\Phi}$$

at a certain location in space and with a time origin chosen to coincide with the peak of the pulse. Φ is a parameter characterizing the amount of chirp.

(a) Sketch the real part of the waveform for $\Phi = 0$ and $\Phi = \sqrt{3}$.

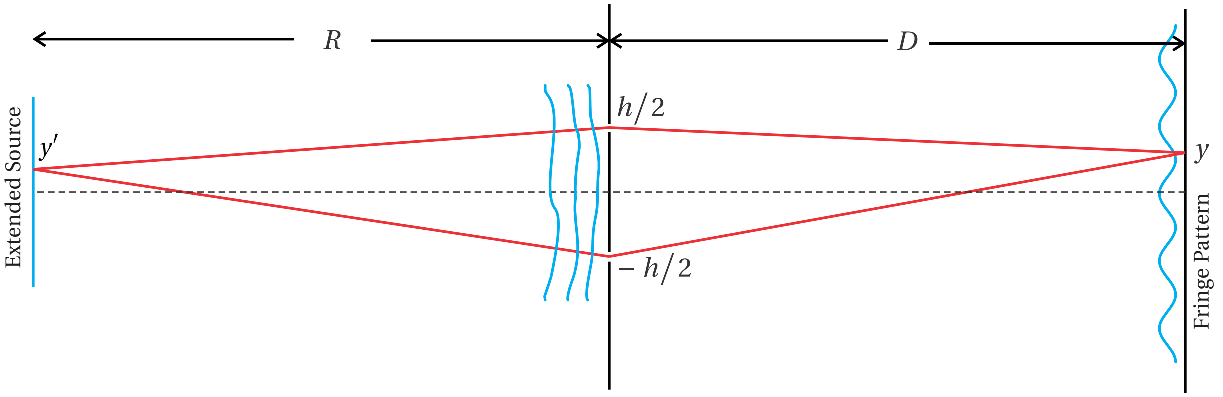
(b) Find $E(\omega)$ and $I(\omega)$. HINT: See integral provided in R44(b).

(c) Compute the degree of coherence $\gamma(\tau)$. How does it depend on Φ?

R47 A diffuse source of light impinges on a Young's double slit (with slit separation h) positioned a distance R from the source. A screen is placed a distance D following the slit. The degree of coherence is given by

$$\gamma(h) \equiv \frac{e^{-i\frac{khy}{D}} \int\limits_{-\infty}^{\infty} I(y')\, e^{-i\frac{khy'}{R}}\, dy'}{\int\limits_{-\infty}^{\infty} I(y')\, dy'}$$

where y (unprimed) is the position on the screen. The source has an



emission distribution with the form $I(y') = \left(I_0/\Delta y'\right) e^{-y'^2/\Delta y'^2}$.

(a) Compute the function $\gamma(h)$. HINT: See integral provided in R44(b).

(b) The intensity on the screen oscillates as a function of y. As h grows wider, the amplitude of oscillations decreases. How wide must the slit separation h become (in terms of R, k, and $\Delta y'$) to reduce the visibility to

$$V \equiv \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \frac{1}{3}$$

(c) Sketch the intensity at the screen $I \propto 1 + \mathrm{Re}\gamma(h)$ when the visibility is 1/3.

Selected Answers

R40: 51.12°.

R41: (b) 1/4, (c) 1/2.

R44: (b) $3.8 \times 10^{-16}\mathrm{J}/\left(\mathrm{cm}^2 \cdot \mathrm{s}^{-1}\right)$.

R46: (b) partial: $E(\omega) = T E_0 e^{-T^2 \frac{(\omega-\omega_0)^2}{2}(1-i\Phi)}$.

Chapter 9

Light as Rays

So far in our study of optics, we have described light in terms of waves, which satisfy Maxwell's equations. However, as you are probably aware, in many situations light can be thought of as *rays* pointing along the direction of wave propagation. A ray picture is useful when one is interested in the macroscopic flow of light energy, but rays fail to reveal fine details, in particular wave and diffraction phenomena. For example, simple ray theory suggests that a lens can focus light down to a point. However, if a beam of light were concentrated onto a true point, the intensity would be infinite! Nevertheless, ray theory is useful for predicting where a focus occurs. It is also useful for describing imaging properties of optical systems (e.g. lenses and mirrors).

Beginning in section 9.3 we study the details of ray theory and the imaging properties of optical systems. First, however, we examine the justification for ray theory starting from Maxwell's equations. In the short-wavelength limit, Maxwell's equations give rise to the *eikonal equation*, which governs the direction of rays in a medium with an index of refraction that varies with position. The German word 'eikonal' comes from the Greek '*εικων*' from which the modern word 'icon' derives. The eikonal equation therefore has a descriptive title since it controls the formation of images. The eikonal equation provides an adequate description as long as the features of interest are large compared to a wavelength.

The eikonal equation describes the direction of ray propagation, even in complicated situations such as desert mirages where air is heated near the ground and has a different index than the air farther from the ground. Rays of light from the sky that initially are directed toward the ground can be bent such that they travel parallel to or even up from the ground, owing to the inhomogeneous refractive index. The eikonal equation can also be used to deduce *Fermat's principle*, which in short says that light travels from point A to point B following a path that takes the minimum time. This principle can be used, for example, to 'derive' Snell's law. Fermat asserted this principle more than a century before Maxwell's equations were known, but it is nice to give justification retroactively using the modern perspective.

Much of this chapter is devoted to the propagation of rays through optical systems composed of lenses and/or curved mirrors in the context of *paraxial ray theory*. The *paraxial approximation* restricts rays to travel *nearly* parallel to the axis of such systems. We consider the effects of three basic optical elements acting on paraxial rays: 1) Free propagation through a distance d in a uniform medium; a ray may move farther away from (or closer to) the *optical axis,* as it travels. 2) Reflection from a curved spherical mirror, which changes a ray's angle with respect to the optical axis. 3) Transmission through a spherical interface between two materials with differing refractive indices. The effects of each of these basic elements on a ray of light can be represented as a 2 × 2 matrix, which can be multiplied together to construct more complex imaging systems (such as a lens or a series of lenses and curved mirrors).

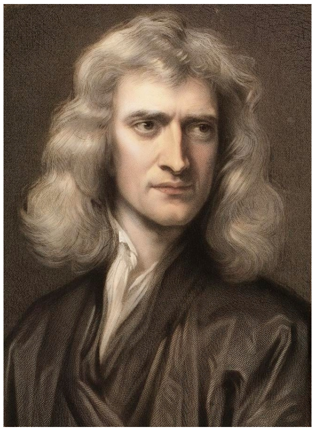
We will study image formation in the context of the paraxial approximation, which in the case of a curved mirror or a *thin* lens gives rise to the familiar formula

$$\frac{1}{f} = \frac{1}{d_\mathrm{o}} + \frac{1}{d_\mathrm{i}} \tag{9.1}$$

Even a complicated multi-element optical system obeys (9.1) if d_o and d_i are measured from *principal planes* rather than the single plane of, for example, a thin lens.

Paraxial ray theory can also be used to study the stability of laser cavities. The formalism predicts whether a ray, after many round trips in the cavity, remains near the optical axis (trapped and therefore stable) or if it drifts endlessly away from the axis of the cavity on successive round trips.

In appendix 9.A we address deviations from the paraxial ray theory known as aberrations. We also comment on ray-tracing techniques, used for designing optical systems that minimize such aberrations.



Sir Isaac Newton (1643–1727, English) was born in Lincolnshire, England three months after the death of his father who was a farmer. Newton spent much of his childhood with his maternal grandmother, after his mother remarried. (Newton did not like his stepfather.) In his teenage years, Newton's mother tried to persuade him to take up farming, but his love for education won out. He became the top-ranked student at his school and was admitted into Trinity College, Cambridge at age 18. Newton was influenced by the works of Descartes, Copernicus, Galileo, and Kepler. Upon graduation four years later, the university closed for two years because of a plague. Newton's return to farm life coincided with a remarkable period when he first developed ideas on calculus, gravitation, and optics. Newton later returned to Cambridge where he spent his extraordinarily prolific career and became the first scientist to be knighted. In optics, Newton advanced the ray theory of light and image formation. He wrote a landmark textbook on the subject. He also showed that 'white' light is comprised of many colors and that the amount of refraction depends on color. He built the first reflecting telescope, which avoids chromatic aberration. Newton advocated against the wave theory of light in favor of his 'corpuscular' theory. (Imagining that Newton foresaw the quantized nature of light energy gives too much credit!) (Wikipedia)

9.1 The Eikonal Equation

For simplicity, consider light consisting of only a single frequency ω. The wave equation (2.13) for an isotropic medium with a real refractive index may be written as

$$\nabla^2 \mathbf{E}(\mathbf{r}, t) + \frac{[n(\mathbf{r})]^2 \omega^2}{c^2} \mathbf{E}(\mathbf{r}, t) = 0 \tag{9.2}$$

where we have already performed the time differentiation on the assumed single-frequency time dependence $e^{-i\omega t}$. Although in chapter 2 we considered solutions to the wave equation in a homogeneous material, the wave equation remains valid when the index of refraction varies throughout space (i.e. if $n(\mathbf{r})$ is an arbitrary function of $\mathbf{r}$). In this case, the usual plane-wave solutions no longer satisfy the wave equation.

As a trial solution for (9.2), we take

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0(\mathbf{r})\, e^{i[k_\mathrm{vac} R(\mathbf{r}) - \omega t]} \tag{9.3}$$

where

$$k_{\text{vac}} = \frac{\omega}{c} = \frac{2\pi}{\lambda_{\text{vac}}} \tag{9.4}$$

Here $R(\mathbf{r})$ is a real scalar function (which depends on position) having the dimension of length. By taking $R(\mathbf{r})$ to be real, there is no absorption or amplification. Even though the trial solution (9.3) looks somewhat like a plane wave,[1] the function $R(\mathbf{r})$ accommodates wavefronts that can be curved or distorted as depicted in Fig. 9.1. At any given instant t, the phase of the curved surfaces described by $R(\mathbf{r}) = \text{constant}$ can be interpreted as wavefronts. The wavefronts travel in the direction for which $R(\mathbf{r})$ varies the fastest. This direction is aligned with $\nabla R(\mathbf{r})$, which lies in the direction perpendicular to surfaces of constant phase.

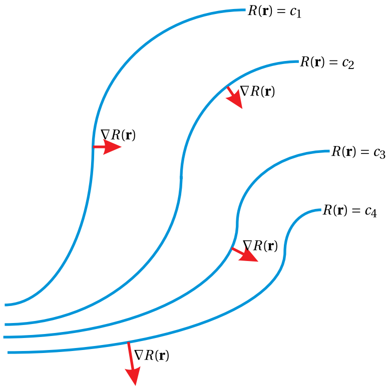


Figure 9.1 Wave fronts (i.e. surfaces of constant phase given by $R(\mathbf{r})$) distributed throughout space in the presence of a spatially inhomogeneous refractive index. The gradient of R gives the direction of travel for a wavefront.

The substitution of the trial solution (9.3) into the wave equation (9.2) gives

$$\frac{1}{k_{\text{vac}}^2}\nabla^2\left[\mathbf{E}_0(\mathbf{r})\, e^{ik_{\text{vac}}R(\mathbf{r})}\right] + [n(\mathbf{r})]^2\,\mathbf{E}_0(\mathbf{r})\, e^{ik_{\text{vac}}R(\mathbf{r})} = 0 \tag{9.5}$$

where we have divided each term by $e^{-i\omega t}$.

Computing the Laplacian in (9.5)

The gradient of the x component of the field is

$$\nabla\left[E_{0x}(\mathbf{r})\, e^{ik_{\text{vac}}R(\mathbf{r})}\right] = [\nabla E_{0x}(\mathbf{r})]\, e^{ik_{\text{vac}}R(\mathbf{r})} + ik_{\text{vac}}E_{0x}(\mathbf{r})\,[\nabla R(\mathbf{r})]\, e^{ik_{\text{vac}}R(\mathbf{r})}$$

The Laplacian of the x component is

$$\begin{aligned}\nabla\cdot\nabla\left[E_{0x}(\mathbf{r})\, e^{ik_{\text{vac}}R(\mathbf{r})}\right] = {} & \{\nabla^2 E_{0x}(\mathbf{r}) - k_{\text{vac}}^2 E_{0x}(\mathbf{r})\,[\nabla R(\mathbf{r})]\cdot[\nabla R(\mathbf{r})] \\ & + ik_{\text{vac}}E_{0x}(\mathbf{r})\left[\nabla^2 R(\mathbf{r})\right] + 2ik_{\text{vac}}\,[\nabla E_{0x}(\mathbf{r})]\cdot[\nabla R(\mathbf{r})]\}\, e^{ik_{\text{vac}}R(\mathbf{r})}\end{aligned}$$

Upon combining the result for each vector component of $\mathbf{E}_0(\mathbf{r})$, the required spatial derivative can be written as

$$\begin{aligned}\nabla^2\left[\mathbf{E}_0(\mathbf{r})\, e^{ik_{\text{vac}}R(\mathbf{r})}\right] = {} & \left(\nabla^2\mathbf{E}_0(\mathbf{r}) - k_{\text{vac}}^2\mathbf{E}_0(\mathbf{r})\,[\nabla R(\mathbf{r})]\cdot[\nabla R(\mathbf{r})] + ik_{\text{vac}}\mathbf{E}_0(\mathbf{r})\left[\nabla^2 R(\mathbf{r})\right]\right. \\ & + 2ik_{\text{vac}}\left\{\hat{\mathbf{x}}\,[\nabla E_{0x}(\mathbf{r})]\cdot[\nabla R(\mathbf{r})] + \hat{\mathbf{y}}\left[\nabla E_{0y}(\mathbf{r})\right]\cdot[\nabla R(\mathbf{r})]\right. \\ & \left.\left. + \;\hat{\mathbf{z}}\,[\nabla E_{0z}(\mathbf{r})]\cdot[\nabla R(\mathbf{r})]\right\}\right) e^{ik_{\text{vac}}R(\mathbf{r})}\end{aligned}$$

After performing the Laplacian and after some rearranging, (9.5) becomes

$$\begin{aligned}\left[\nabla R(\mathbf{r})\cdot\nabla R(\mathbf{r}) - [n(\mathbf{r})]^2\right]\mathbf{E}_0(\mathbf{r}) = {} & \frac{\nabla^2\mathbf{E}_0(\mathbf{r})}{k_{\text{vac}}^2} + \frac{i}{k_{\text{vac}}}\nabla^2 R(\mathbf{r}) + \frac{2i}{k_{\text{vac}}}\hat{\mathbf{x}}\nabla E_{0x}(\mathbf{r})\cdot\nabla R(\mathbf{r}) \\ & + \frac{2i}{k_{\text{vac}}}\hat{\mathbf{y}}\nabla E_{0y}(\mathbf{r})\cdot\nabla R(\mathbf{r}) + \hat{\mathbf{z}}\frac{2i}{k_{\text{vac}}}\nabla E_{0z}(\mathbf{r})\cdot\nabla R(\mathbf{r})\end{aligned} \tag{9.6}$$

---

[1] If the index is spatially independent (i.e. $n(\mathbf{r}) \to n$), then (9.3) reduces to the usual plane-wave solution of the wave equation. In this case, we have $R(\mathbf{r}) = \mathbf{k}\cdot\mathbf{r}/k_{\text{vac}}$ and the field amplitude becomes constant (i.e. $\mathbf{E}_0(\mathbf{r}) \to \mathbf{E}_0$).

Don't be afraid; at this point we are ready to make an important approximation. We take the limit of a very short wavelength (i.e. $1/k_{\text{vac}} = \lambda_{\text{vac}}/2\pi \to 0$), and the entire right-hand side of (9.6) vanishes. (Thank goodness!) With it we lose the effects of diffraction so that only macroscopic features are relevant. The equation also knows nothing about surface reflections at abrupt index changes.

Our wave equation has been simplified to

$$[\nabla R(\mathbf{r})] \cdot [\nabla R(\mathbf{r})] = [n(\mathbf{r})]^2 \tag{9.7}$$

Written another way, this equation is

$$\nabla R(\mathbf{r}) = n(\mathbf{r})\,\hat{\mathbf{s}}(\mathbf{r}) \tag{9.8}$$

where $\hat{\mathbf{s}}$ is a unit vector pointing in the direction $\nabla R(\mathbf{r})$, the direction normal to wavefront surfaces. Equation (9.8) is called the *eikonal equation*.[2]

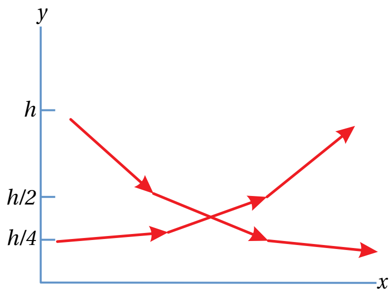


Figure 9.2 Depiction of possible light ray paths in a region with varying index.

Example 9.1

Suppose that a region of air above the desert on a hot day has an index of refraction that varies with height y according to $n(y) = n_0\sqrt{1 + y^2/h^2}$. Verify that $R(x, y) = n_0\left(x \pm y^2/2h\right)$ is a solution to the eikonal equation. (See problem P9.1 for a more general solution.)

Solution: The gradient of our trial solution gives

$$\nabla R(x, y) = n_0\left(\hat{\mathbf{x}} \pm \hat{\mathbf{y}}\, y/h\right)$$

Substituting this into (9.7) gives

$$\nabla R \cdot \nabla R = n_0\left(\hat{\mathbf{x}} \pm \hat{\mathbf{y}}\, y/h\right) \cdot n_0\left(\hat{\mathbf{x}} \pm \hat{\mathbf{y}}\, y/h\right) = n_0^2\left(1 + y^2/h^2\right) = \left[n(y)\right]^2$$

which confirms that it is a solution. The direction of light propagation is

$$\hat{\mathbf{s}}(y) \equiv \frac{\nabla R}{|\nabla R|} = \frac{n_0\left(\hat{\mathbf{x}} \pm \hat{\mathbf{y}}\, y/h\right)}{n_0\sqrt{1 + y^2/h^2}} = \frac{\hat{\mathbf{x}} \pm \hat{\mathbf{y}}\, y/h}{\sqrt{1 + y^2/h^2}}$$

Computed at various heights, the direction for rays turns out to be

$$\hat{\mathbf{s}}(h) = \frac{\hat{\mathbf{x}} \pm \hat{\mathbf{y}}}{\sqrt{2}} \qquad \hat{\mathbf{s}}(h/2) = \frac{\hat{\mathbf{x}} \pm \hat{\mathbf{y}}/2}{\sqrt{5/4}} \qquad \hat{\mathbf{s}}(h/4) = \frac{\hat{\mathbf{x}} \pm \hat{\mathbf{y}}/4}{\sqrt{17/16}}$$

These are represented in Fig. 9.2. In a desert mirage, light from the sky can appear to come from a lower position. We can determine a path for the rays by setting dy/dx equal to the slope of $\hat{\mathbf{s}}$:

$$\frac{dy}{dx} = \pm\frac{y}{h} \quad \Rightarrow \quad y = y_0 e^{\pm(x-x_0)/h}$$

---

[2] M. Born and E. Wolf, *Principles of Optics*, 7th ed., Sect. 3.1.1 (Cambridge University Press, 1999).

It can be shown that the Poynting vector is directed along $\hat{\mathbf{s}}$ (see P9.2). In other words, the direction of $\hat{\mathbf{s}}$ specifies the direction of energy flow. The unit vector $\hat{\mathbf{s}}$ at each location in space points perpendicular to the wavefronts and indicates the direction that the waves travel as seen in Fig. 9.1. A collection of vectors $\hat{\mathbf{s}}$ distributed throughout space are called *rays*.

In retrospect, we might have jumped straight to (9.8) without going through the above derivation. After all, we know that each part of a wavefront advances in the direction of its gradient $\nabla R(\mathbf{r})$ (i.e. in the direction that $R(\mathbf{r})$ varies most rapidly). We also know that each part of a wavefront defined by $R(\mathbf{r}) = \text{constant}$ travels at speed $c/n(\mathbf{r})$. The slower a given part of the wavefront advances, the more rapidly $R(\mathbf{r})$ changes with position $\mathbf{r}$ and the closer the contours of constant phase. It follows that $\nabla R(\mathbf{r})$ must be proportional to $n(\mathbf{r})$ since $\nabla R(\mathbf{r})$ denotes the rate of change in $R(\mathbf{r})$.

9.2 Fermat's Principle

As we have seen, the eikonal equation (9.8) governs the path that rays follow as they traverse a region of space, where the index varies with position. Another way of deducing the correct path of rays is via Fermat's principle.[3] Fermat's principle says that if a ray happens to travel through both points A and B, it will follow a path between them that takes the least time.

Pierre de Fermat (1601–1665, French) was born in Beaumont-de-Lomagne, France to a wealthy merchant family. He attended the University of Toulouse before moving to Bordeaux in the late 1620s where Fermat distinguished himself as a mathematician. Fermat was proficient in many languages and went on to obtain a law degree in 1631 from the University of Orleans. He continued his study of mathematics as a hobby throughout his life. He corresponded with a number of notable mathematicians, and through his letters made notable contributions to analytic geometry, probability theory, and number theory. He was often quite secretive about the methods used to obtain his results. Mathematicians suspect that Fermat didn't actually prove his famous last theorem, which was not able to be verified until the 1990's. Fermat was the first to assert that the path taken by a beam of light is the one that can be traveled in the least amount of time. (Wikipedia)

Derivation of Fermat's Principle from the Eikonal Equation

We begin by taking the curl of (9.8) to obtain[4]

$$\nabla \times [n(\mathbf{r})\,\hat{\mathbf{s}}(\mathbf{r})] = \nabla \times [\nabla R(\mathbf{r})] = 0 \tag{9.9}$$

This can be integrated over an open surface of area A to give

$$\int_A \nabla \times [n(\mathbf{r})\,\hat{\mathbf{s}}(\mathbf{r})]\, da = \oint_C n(\mathbf{r})\,\hat{\mathbf{s}}(\mathbf{r}) \cdot d\boldsymbol{\ell} = 0 \tag{9.10}$$

We have applied Stokes' theorem (0.12) to convert the area integral into a path integral around the perimeter contour C.

Equation (9.10) states that the integration of $n\hat{\mathbf{s}} \cdot d\boldsymbol{\ell}$ around a closed loop is always zero. If we consider a closed loop comprised of a path from point A to point B and then a different path from B back to A again, the integrals for the two legs always cancel, even while holding one path fixed while varying the other. This means

$$\int_A^B n\hat{\mathbf{s}} \cdot d\boldsymbol{\ell} \quad \text{is independent of path from A to B.} \tag{9.11}$$

Now consider a path from A to B that is parallel to $\hat{\mathbf{s}}$, as depicted in Fig. 9.3. In this case, the cosine in the dot product is always one. If we choose some other

---

[3] M. Born and E. Wolf, *Principles of Optics*, 7th ed., Sect. 3.3.2 (Cambridge University Press, 1999).

[4] The curl of a gradient is identically zero for any function.

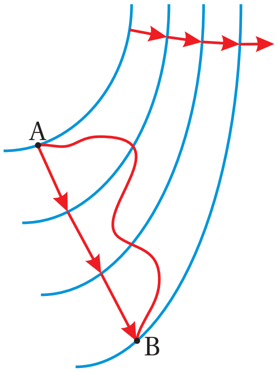


Figure 9.3 A ray of light leaving point A arriving at B.

path that connects A and B, the cosine associated with the dot product is less than one at most points along that path, whereas the result of the integral is the same. Therefore, if we artificially remove the dot product from the integral (i.e. exclude the cosine factor), the result of the integral will exceed the true value unless the path chosen follows the direction of $\hat{\mathbf{s}}$ (i.e. the path that corresponds to the one that light rays actually follow).

In mathematical form, this argument can be expressed as

$$\int_A^B n\hat{\mathbf{s}}\cdot d\ell = \min\left\{\int_A^B n\,d\ell\right\} \tag{9.12}$$

The integral on the right is called the *optical path length* (OPL) between points A and B:

$$OPL|_A^B \equiv \int_A^B n\,d\ell \tag{9.13}$$

The conclusion is that the true path that light follows between two points (i.e. the one that stays parallel to $\hat{\mathbf{s}}$) is the one with the shortest optical path length. The index n may vary with position and therefore can be different for each of the incremental distances $d\ell$.

Fermat's principle is usually stated in terms of the time it takes light to travel between points. The travel time Δt depends not only on the path taken by the light but also on the velocity of the light $v(\mathbf{r})$, which varies spatially with the refractive index:

$$\Delta t|_A^B = \int_A^B \frac{d\ell}{v(\mathbf{r})} = \int_A^B \frac{d\ell}{c/n(\mathbf{r})} = \frac{OPL|_A^B}{c} \tag{9.14}$$

To find the correct path for the light ray that leaves point A and crosses point B, we need only minimize the optical path length between the two points (proportional to travel time). The optical path length is not the actual distance that the light travels; it is proportional to the number of wavelengths that fit into that distance (see (2.24)). Thus, as the wavelength shortens due to a higher index of refraction, the optical path length increases. The correct ray traveling from A to B does not necessarily follow a straight line but can follow a complicated curve according to how the index varies.

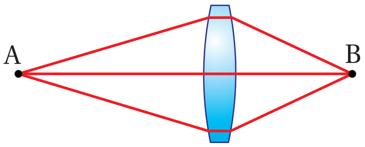


Figure 9.4 Rays of light leaving point A with the same optical path length to B.

An imaging situation occurs when many paths from point A to point B have the same optical path length. An example of this occurs when a lens causes an image to form. In this case all rays leaving point A (on an object) and traveling through the system to point B (on the image) experience equal optical path lengths. Although the ray traveling through the center of the lens depicted in Fig. 9.4 has a shorter geometric path length, it goes through more material so that the optical path length is the same as for the outer rays.

To summarize Fermat's principle,[5] of the many rays that might emanate from a point A, the ray that crosses a second point B is the one that follows the shortest optical path length. If many rays tie for having the shortest optical path, we say that an *image* of point A forms at point B.

---

[5] The minimization of (9.14) does not give the correct path in anisotropic media such as crystals where n depends on the direction of a ray as well as on its location (see P9.5).

Example 9.2

Use Fermat's principle to derive Snell's law.

Solution: Consider the many rays of light that leave point A seen in Fig. 9.5. Only one of the rays passes through point B. Within each medium we expect the light to travel in a straight line since the index is uniform. However, at the boundary we must allow for bending since the index changes.

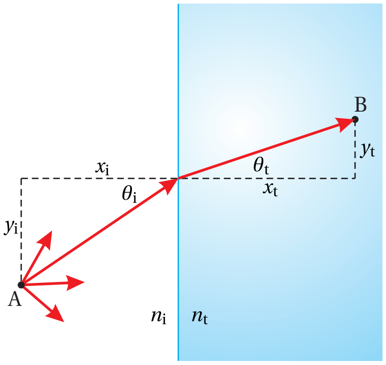


Figure 9.5 Rays of light leaving point A; not all of them will traverse point B.

The optical path length between points A and B may be written

$$OPL = n_i\sqrt{x_i^2 + y_i^2} + n_t\sqrt{x_t^2 + y_t^2} \tag{9.15}$$

We need to minimize this optical path length to find the correct one according to Fermat's principle.

Since points A and B are fixed, we may regard x_i and x_t as constants. The distances y_i and y_t are not constants although the combination

$$y_{\text{tot}} = y_i + y_t \tag{9.16}$$

is constant. Thus, we may rewrite (9.15) as

$$OPL(y_i) = n_i\sqrt{x_i^2 + y_i^2} + n_t\sqrt{x_t^2 + (y_{\text{tot}} - y_i)^2} \tag{9.17}$$

where the only variable is y_i.

We now minimize the optical path length by taking the derivative and setting it equal to zero:

$$\frac{d(OPL)}{dy_i} = n_i\frac{y_i}{\sqrt{x_i^2 + y_i^2}} + n_t\frac{-(y_{\text{tot}} - y_i)}{\sqrt{x_t^2 + (y_{\text{tot}} - y_i)^2}} = 0 \tag{9.18}$$

Notice that

$$\sin\theta_i = \frac{y_i}{\sqrt{x_i^2 + y_i^2}} \quad \text{and} \quad \sin\theta_t = \frac{y_t}{\sqrt{x_t^2 + y_t^2}} \tag{9.19}$$

With these substitutions, (9.18) reduces to

$$n_i \sin\theta_i = n_t \sin\theta_t \tag{9.20}$$

which is the familiar Snell's law.

Example 9.3

Use Fermat's principle to derive the equation of curvature for a reflective surface that causes all rays leaving one point to image to another. Do the calculation in two dimensions rather than in three.[6]

Solution: We adopt the convention that the origin is halfway between the points, which are separated by a distance $2a$, as shown in Fig. 9.6. If the points are to image to each other, Fermat's principle requires that the total path length be a constant; call it b. If the total path from the 'object' point to the 'image' point includes a reflection from an arbitrary point (x, y), we may write constant path length as

$$\sqrt{(x+a)^2+y^2}+\sqrt{(x-a)^2+y^2}=b \tag{9.21}$$

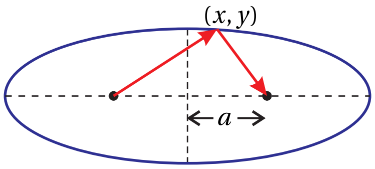


Figure 9.6

To get (9.21) into a more recognizable form, we isolate the first square root and square both sides of the equation, which gives

$$(x+a)^2+y^2=b^2+(x-a)^2+y^2-2b\sqrt{(x-a)^2+y^2}$$

After squaring the two binomial terms, some nice cancelations occur, and we get

$$4ax-b^2=-2b\sqrt{(x-a)^2+y^2}$$

which we square again to obtain

$$16a^2x^2-4ab^2x+b^4=4b^2\left(x^2-2ax+a^2+y^2\right)$$

After some cancellations and regrouping this becomes

$$\left(16a^2-4b^2\right)x^2-4b^2y^2=4a^2b^2-b^4$$

Finally, we divide both sides by the term on the right to obtain the (hopefully) familiar form of an ellipse

$$\frac{x^2}{\left(\frac{b^2}{4}\right)}+\frac{y^2}{\left(\frac{b^2}{4}-a^2\right)}=1 \tag{9.22}$$

9.3 Paraxial Rays and ABCD Matrices

We now turn our attention to the effects of curved mirrors and lenses on rays of light. Keep in mind that when describing light as a collection of rays rather than as waves, the results can only describe features that are macroscopic compared to a wavelength. The rays of light at each location in space describe approximately the direction of travel of the wavefronts at that location. Since the wavelength of visible light is extraordinarily small compared to the macroscopic features that we perceive in our day-to-day world, the ray approximation is often a very good one. This is the reason that ray optics was developed long before light was understood as a wave.

---

[6] This configuration is used to direct flash lamp energy into a laser amplifier rod. One 'point' in Fig. 9.6 represents the end of an amplifier rod while the other represents the end of a thin flash-lamp tube.

We consider ray theory within the *paraxial approximation*, meaning that we restrict our attention to rays that are near and almost parallel to an *optical axis* of a system, say the z-axis. It is within this approximation that the familiar imaging properties of lenses occur. An image occurs when all rays from a *point* on an *object* converge to a corresponding *point* on what is referred to as the *image*. To the extent that the paraxial approximation is violated, the clarity of an image can suffer, and we say that there are *aberrations* present. The field of optical engineering is often concerned with minimizing aberrations in cases where the paraxial approximation is not strictly followed. This is done so that, for example, a camera can take pictures of objects that occupy a fairly wide angular field of view, where rays violate the paraxial approximation. Optical systems are typically engineered using the science of *ray tracing*, which is described briefly in section 9.A.

As we develop paraxial ray theory, we should remember that rays impinging on devices such as lenses or curved mirrors should strike the optical component at near normal incidence. To quantify this statement, the paraxial approximation is valid to the extent that

$$\sin\theta \cong \theta \tag{9.23}$$

is a good approximation, and similarly

$$\tan\theta \cong \theta \tag{9.24}$$

Here, the angle θ (in radians) represents the angle that a particular ray makes with respect to the optical axis. There is an important mathematical reason for this approximation. The sine is a nonlinear function, but at small angles it is approximately linear and can be represented by its argument. This linearity greatly simplifies the analysis since it reduces the problem to linear algebra. Conveniently, we will be able to keep track of imaging effects with a 2×2 matrix formalism.

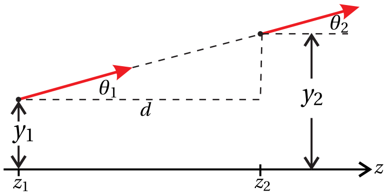


Figure 9.7 The behavior of a ray as light traverses a distance d.

Consider a ray propagating in the y–z plane where the optical axis is in the z-direction. Let us specify a ray at position z_1 by two coordinates: the displacement from the axis y_1 and the orientation angle θ_1 (see Fig. 9.7). If the index is uniform everywhere, the ray travels along a straight path. It is straightforward to predict the coordinates of the same ray downstream, say at z_2. First, since the ray continues in the same direction, we have

$$\theta_2 = \theta_1 \tag{9.25}$$

By referring to Fig. 9.7 we can write y_2 in terms of y_1 and θ_1:

$$y_2 = y_1 + d\tan\theta_1 \tag{9.26}$$

where $d \equiv z_2 - z_1$. Equation (9.26) is nonlinear in θ_1, but within the paraxial approximation it becomes simply

$$y_2 = y_1 + \theta_1 d \tag{9.27}$$

Equations (9.25) and (9.27) describe a linear transformation, which in matrix notation may be written as

ABCD matrix for propagation through a distance d

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \tag{9.28}$$

Here, the vectors in this equation specify the essential information about the ray before and after traversing the distance d, and the matrix describes the effect of traversing the distance. This type of matrix is called an *ABCD matrix*;[7] sometimes physicists are not very inventive with names.

Example 9.4

Let the distance d be subdivided into two distances, a and b, such that $d = a + b$. Show that an application of the ABCD matrix for distance a followed by an application of the ABCD matrix for b renders same result as an application of the ABCD matrix for distance d.

Solution: Individually, the effects of propagation through a and through b are

$$\begin{bmatrix} y_{\text{mid}} \\ \theta_{\text{mid}} \end{bmatrix} = \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{\text{mid}} \\ \theta_{\text{mid}} \end{bmatrix} \tag{9.29}$$

where the subscript “mid” refers to the ray in the middle position after traversing the distance a. If we combine the equations, we get

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \tag{9.30}$$

which is in agreement with (9.28) since the ABCD matrix for both displacements is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & a+b \\ 0 & 1 \end{bmatrix} \tag{9.31}$$

9.4 Reflection and Refraction at Curved Surfaces

We next consider the effect of *reflection* from a *spherical surface* as depicted in Fig. 9.8. We consider only the act of reflection without considering propagation before or after the reflection takes place. Thus, the incident and reflected rays in the figure are symbolic only of the direction of propagation before and after reflection; they do not indicate any amount of travel. We immediately write

$$y_2 = y_1 \tag{9.32}$$

since the ray has no chance to go anywhere.

---

[7] P. W. Milonni and J. H. Eberly, *Lasers*, Sect. 14.2 (New York: Wiley, 1988).

We adopt the widely used convention that, upon reflection, the positive z-direction is reoriented so that we consider the rays still to travel in the positive z sense. An easy way to remember this is that the positive z direction is always taken to be down stream of where the light is headed. Notice that in Fig. 9.8, the reflected ray approaches the z-axis. In this case θ_2 is a negative angle (as opposed to θ_1 which is drawn as a positive angle) and is equal to

$$\theta_2 = -(\theta_1 + 2\theta_\mathrm{i}) \tag{9.33}$$

where θ_i is the angle of incidence with respect to the normal to the spherical mirror surface. By the law of reflection, the incident and reflected ray both occur at an angle θ_i, referenced to the surface normal. The surface normal points towards the center of curvature of the mirror surface, which we assume is on the z-axis a distance R away. By convention, the radius of curvature R is positive if the mirror surface is *concave* and negative if the mirror surface is *convex*.

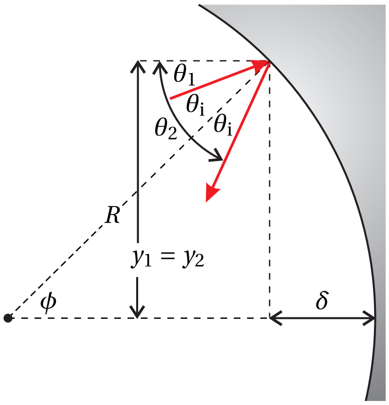


Figure 9.8 A ray depicted in the act of reflection from a spherical surface.

Elimination of θ_i from (9.33) in favor of θ_1 and y_1

By inspection of Fig. 9.8 we can write

$$\frac{y_1}{R} = \sin\phi \cong \phi \tag{9.34}$$

where we have applied the paraxial approximation (9.23). (The angles in Fig. 9.8 are exaggerated. In fact, when ϕ is small enough for (9.34) to hold, we may also neglect the small distance δ.) By inspection of the geometry, we also have

$$\phi = \theta_1 + \theta_\mathrm{i} \tag{9.35}$$

and when this is combined with (9.34), we get

$$\theta_\mathrm{i} = \frac{y_1}{R} - \theta_1 \tag{9.36}$$

With this we are able to put (9.33) into a useful linear form:

$$\theta_2 = -\frac{2}{R} y_1 + \theta_1 \tag{9.37}$$

Equations (9.32) and (9.37) describe a linear transformation that can be concisely formulated as

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -2/R & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \tag{9.38}$$

ABCD matrix for a curved mirror

The ABCD matrix in this transformation describes the act of reflection from a concave mirror with radius of curvature R. As noted, the radius R is negative when the mirror is convex.

The final basic element that we shall consider is a *spherical interface* between two materials with indices n_i and n_t (see Fig. 9.9). This has an effect similar to that of the curved mirror, which changes the direction of a ray without altering

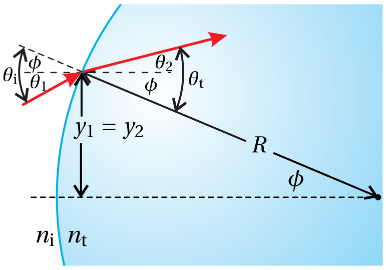


Figure 9.9 A ray depicted in the act of transmission at a curved material interface.

its distance y_1 from the optical axis. Please note that here the radius of curvature is considered to be positive for a *convex surface* (opposite convention from that of the mirror). In this way, if the lower index is on the left, a positive radius R for the interface tends to deflect rays towards the axis just as a positive radius for a mirror does. Again, we are interested only in the act of transmission without any travel before or after the interface. As before, (9.32) applies (i.e. $y_2 = y_1$).

At the interface, the rays obey Snell's law, which in the paraxial approximations is written

$$n_i\theta_i = n_t\theta_t \tag{9.39}$$

The angles θ_i and θ_t are referenced from the surface normal, as seen in Fig. 9.9.

Substituting θ_1, θ_2 and y_1 into Snell's Law

By inspection of Fig. 9.9, we have

$$\theta_i = \theta_1 + \phi \tag{9.40}$$

and

$$\theta_t = \theta_2 + \phi \tag{9.41}$$

where ϕ is the angle that the surface normal makes with the z-axis. As before (see (9.34)), within the paraxial approximation we may write

$$\phi \cong y_1/R$$

When this is used in (9.40) and (9.41), which are substituted into (9.39), Snell's law becomes

$$\theta_2 = \left(\frac{n_i}{n_t} - 1\right)\frac{y_1}{R} + \frac{n_i}{n_t}\theta_1 \tag{9.42}$$

The compact matrix form of (9.32) and (9.42) is written

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ (n_i/n_t - 1)/R & n_i/n_t \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix}$$

(ABCD matrix for a curved interface) (9.43)

Distance within a material, excluding interfaces

$$\begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}$$

Window, starting and stopping in air

$$\begin{bmatrix} 1 & d/n \\ 0 & 1 \end{bmatrix}$$

Thin lens or Mirror

$$\begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix}$$

Thin Lens: $\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$
Mirror: $\frac{1}{f} = \frac{2}{R}$

Thick lens

$$\begin{bmatrix} 1-\frac{d}{R_1}\left(1-\frac{1}{n}\right) & \frac{d}{n} \\ -(n-1)\left(\frac{1}{R_1}-\frac{1}{R_2}\right)+\frac{d}{R_1R_2}\left(2-n-\frac{1}{n}\right) & 1+\frac{d}{R_2}\left(1-\frac{1}{n}\right) \end{bmatrix}$$

Table 9.1 Summary of ABCD matrices for common optical elements.

9.5 ABCD Matrices for Combined Optical Elements

To summarize the previous two sections, we have developed ABCD matrices for three basic elements: 1) propagation through a region of uniform index (9.28), 2) reflection from a curved mirror (9.38), and 3) transmission through a curved interface between regions with different indices (9.43). All other ABCD matrices that we will use are composites of these three. For example, one can construct the ABCD matrix for a *lens* by using two matrices like those in (9.43) to represent the entering and exiting surfaces of the lens. A distance matrix (9.28) can be inserted to account for the thickness of the lens. It is left as an exercise to derive the ABCD matrix for a thick lens (see P9.7).

Example 9.5

Derive the ABCD matrix for a *thin lens*, where the thickness between the two lens surfaces is ignored. (See P 9.7 for the more general case of a thick lens.)

Solution: A thin lens is depicted in Fig. 9.10. R_1 is the radius of curvature for the first surface (which is positive if convex as drawn), and R_2 is the radius of curvature for the second surface (which is negative as drawn).

We take the index outside of the lens to be unity while that of the lens material to be n. We apply the ABCD matrix (9.43) in sequence, once for entering the lens and once for exiting:

$$\begin{aligned}\begin{bmatrix} A & B \\ C & D \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ \frac{1}{R_2}(n-1) & n \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{1}{R_1}\left(\frac{1}{n}-1\right) & \frac{1}{n} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ -(n-1)\left(\frac{1}{R_1}-\frac{1}{R_2}\right) & 1 \end{bmatrix}\end{aligned} \qquad (9.44)$$

ABCD matrix for a thin lens

The matrix for the first interface is written on the right, where it operates first on an incoming ray vector. In this case, $n_i = 1$ and $n_t = n$. The matrix for the second surface is written on the left so that it operates afterwards. For the second surface, $n_i = n$ and $n_t = 1$.

Notice the close similarity between the ABCD matrix for a thin lens (9.44) and the ABCD matrix for a curved mirror (9.38). The ABCD matrix for either the thin lens or the mirror can be written as

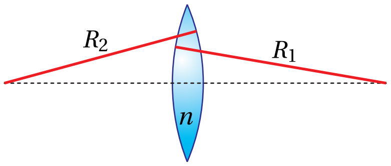


Figure 9.10 Thin lens.

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix} \qquad (9.45)$$

where in the case of the thin lens the *focal length* f is given by *lens maker's formula*

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1}-\frac{1}{R_2}\right) \qquad \text{(focal length of thin lens)} \quad (9.46)$$

and in the case of a curved mirror, the focal length is

$$f = R/2 \qquad \text{(focal length for a curved mirror)} \quad (9.47)$$

The reason for calling f the focal length will become apparent later. Table 9.1 gives a summary of ABCD matrices of common optical elements.

Example 9.6

Derive the ABCD matrix for a window with thickness d and index n.

Solution: We can again take advantage of the ABCD matrix for a curved interface (9.43), only with $R_1 = \infty$ and $R_2 = \infty$ to provide flat surfaces. We take the index outside of the window to be unity and the index inside the window to be n. We use the ABCD matrix (9.43) twice, once for each interface, sandwiching matrix (9.31), which endows the window with thickness:

$$\begin{aligned}\begin{bmatrix} A & B \\ C & D \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ 0 & n \end{bmatrix}\begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{n} \end{bmatrix} \\ &= \begin{bmatrix} 1 & d/n \\ 0 & 1 \end{bmatrix} \qquad \text{(window)}\end{aligned} \tag{9.48}$$

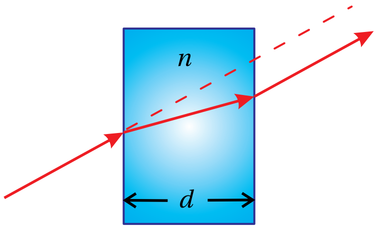


Figure 9.11 Window.

As far as rays are concerned, a window is effectively shorter to traverse than free space.[8] Fig. 9.11 illustrates why this is the case. The displacement of the exiting ray is not as great as it would have been without the window. The window impedes the rate at which the ray can move away from or toward the optical axis.

Example 9.7

Find ray $\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix}$ that results when $\begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix}$ propagates through a distance a, reflects from a mirror of radius R, and then propagates through a distance b. See Fig. 9.12.

Solution: The final ray in terms of the initial one is computed as follows:

$$\begin{aligned}\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} &= \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -2/R & 1 \end{bmatrix}\begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \\ &= \begin{bmatrix} 1-2b/R & a+b-2ab/R \\ -2/R & 1-2a/R \end{bmatrix}\begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \\ &= \begin{bmatrix} (1-2b/R)\,y_1 + (a+b-2ab/R)\,\theta_1 \\ (-2/R)\,y_1 + (1-2a/R)\,\theta_1 \end{bmatrix}\end{aligned} \tag{9.49}$$

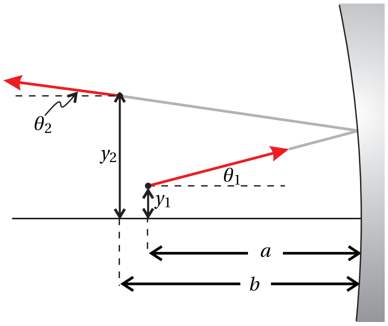


Figure 9.12 A ray that travels through a distance a, reflects from a mirror, and then travels through a distance b.

As always, the ordering of the matrices is important. The first effect that the ray experiences is represented by the matrix on the right, which is in the position that operates first on $\begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix}$.

We have derived our basic ABCD matrices for rays traveling in the y–z plane, as suggested in Figs. 9.7–9.12. This may have given the impression that it is necessary to work within a plane that contains the optical axis (i.e. the z-axis in our case). However, within the paraxial approximation, the ABCD matrices are valid for rays that become displaced simultaneously in both the x and y dimensions during propagating along z.

As we demonstrate below, the behavior of rays functions independently in the x and y dimensions. If desired, one can write a ray vector for each dimension, namely $\begin{bmatrix} x \\ \theta_x \end{bmatrix}$ and $\begin{bmatrix} y \\ \theta_y \end{bmatrix}$. Moreover, the identical matrices, for example any in table 9.1, are used for either dimension. Figs. 9.7–9.12 therefore represent projections of rays onto the y–z plane. To complete the story, one can imagine corresponding figures representing the projection of the rays onto the x–z plane.

---

[8] In contrast, the optical path length OPL is effectively longer than free space by the factor n.

Independence of Rays in the x and y Dimensions

Imagine a ray contained within a plane that is parallel to the y–z plane but for which $x > 0$. One might be concerned that when the ray meets, for example, a spherically concave mirror, the radius of curvature *in the perspective of the y–z dimension* might be different for $x > 0$ than for $x = 0$ (at the center of the mirror). This concern is actually quite legitimate and is the source of what is known as spherical aberration. Nevertheless, in the paraxial approximation the intersection with the curved mirror of all planes that are parallel to the optical axis gives the same curve.

To see why this is so, consider the curvature of the mirror in Fig. 9.8. As we move away from the mirror center (in the x or y-dimension or some combination thereof), the mirror curves to the left by the amount

$$\delta = R - R\cos\phi \tag{9.50}$$

In the paraxial approximation, we have $\cos\phi \cong 1 - \phi^2/2$. And since in this approximation we may also write $\phi \cong \sqrt{x^2+y^2}\Big/R$, (9.50) becomes

$$\delta \cong \frac{x^2}{2R} + \frac{y^2}{2R} \tag{9.51}$$

In the paraxial approximation, we see that the curve of the mirror is parabolic, and therefore separable between the x and y dimensions. That is, the curvature in the x-dimension (i.e. $\partial\delta/\partial x = x/R$) is independent of y, and the curvature in the y-dimension (i.e. $\partial\delta/\partial y = y/R$) is independent of x. A similar argument can be made for a spherical interface between two media.

When many ABCD matrices are multiplied together to represent, for example, a multi-element lens system, remarkably the resulting ABCD matrix has the following property:

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = 1 \tag{9.52}$$

The determinant of the ABCD matrix is one so long as we *begin and end in the same index of refraction.*[9] Notice that the determinants of all of the matrices in table 9.1 are one. Moreover, matrices constructed of matrices whose determinants are one are guaranteed also to have determinants equal to one.

9.6 Image Formation

Consider Example 9.7 where a ray travels a distance a, reflects from a curved mirror, and then travels a distance b. From (9.49), the ABCD matrix for the overall process is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 - b/f & a + b - ab/f \\ -1/f & 1 - a/f \end{bmatrix} \tag{9.53}$$

Galileo di Vincenzo Bonaiuti de' Galilei (1564–1642, Italian) was born in Pisa, Italy, the son of a musician. Galileo enrolled in the University of Pisa with the intent to study medicine but soon became diverted into mathematics. He served three years as chair of mathematics in Pisa beginning in 1589 and then moved to the University of Padua where he taught geometry, mechanics, and astronomy for two decades. While Galileo did not invent the telescope, he improved the design considerably. With it, he discovered four moons of Jupiter and was the first to observe sunspots and mountains and valleys on the Moon. Galileo also was the first to document the phases of Venus, similar to the phases of the moon. He used these observations to argue in favor of the Copernican model of the solar system, but this conflicted with the prevailing views of the Catholic Church at the time. He was placed under house arrest and forbidden to publish of any of his works. While under house arrest, he wrote much on kinematics and other principles of physics and is considered to be the father of modern physics. Galileo attempted to measure the speed of light by observing an assistant uncover a lantern on a distant hill in response to a light signal. He concluded that light is "very swift" if not instantaneous. (Wikipedia) where by (9.47) we have replaced $2/R$ with $1/f$. Because of the similarity between the behavior of a curved mirror and a thin lens, the above expression can also represent a ray traveling a distance a, traversing a thin lens with focal length f, and then traveling a distance b. The only difference is that, in the case of the thin lens, f is given by lens maker's formula (9.46).

---

[9] The determinant of (9.43) is not one since it starts and ends with different indices of refraction. However, when this matrix is used in succession to form a lens, the resulting matrix has determinant equal to one.

As is well known, it is possible to form an image with either a curved mirror or a lens. Suppose that the initial ray is *one of many rays* that leaves a particular point on an object positioned $a = d_o$ before the mirror (or lens). In order for an image to occur at $d_i = b$, it is essential that all rays leaving the particular point on the object converge to a corresponding point on the image. That is, we want rays leaving the point y_1 on the object (which may take on a range of angles θ_1) all to converge to a single point y_2 at the image. In the following equation we need y_2 to be independent of θ_1:

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} Ay_1 + B\theta_1 \\ Cy_1 + D\theta_1 \end{bmatrix} \tag{9.54}$$

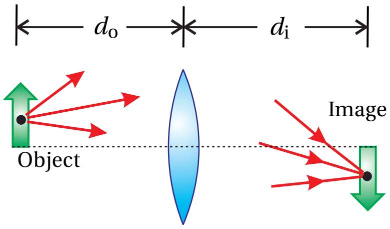


Figure 9.13 Image formation by a thin lens.

The condition for image formation is therefore

$$B = 0 \qquad \text{(condition for image formation)} \tag{9.55}$$

When this condition is applied to (9.53), we obtain

$$d_o + d_i - \frac{d_o d_i}{f} = 0 \Rightarrow \frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i} \tag{9.56}$$

which is the familiar imaging formula (9.1). When the object is infinitely far away (i.e. $d_o \to \infty$), the image appears at $d_i \to f$. This gives a physical interpretation to the *focal length* f, as we have been calling it. Please note that d_o and d_i can each be either positive (*real* as depicted in Fig. 9.13) or negative (*virtual* meaning a screen cannot be inserted to display the image).

The *magnification* of the image is found by comparing the size of y_2 to y_1. From (9.53)–(9.56), the magnification is found to be

$$M \equiv \frac{y_2}{y_1} = A = 1 - \frac{d_i}{f} = -\frac{d_i}{d_o} \tag{9.57}$$

The negative sign indicates that for positive distances d_o and d_i the image is inverted.

> **Example 9.8**
>
> Beginning students are often taught to draw *ray diagrams* such as the one in Fig. 9.14, which shows a *real image* formed by a thin lens. Several key rays aid in a graphic prediction of the location and size of the image. Use ABCD-matrix analysis to describe the effect of the lens on the three rays drawn.

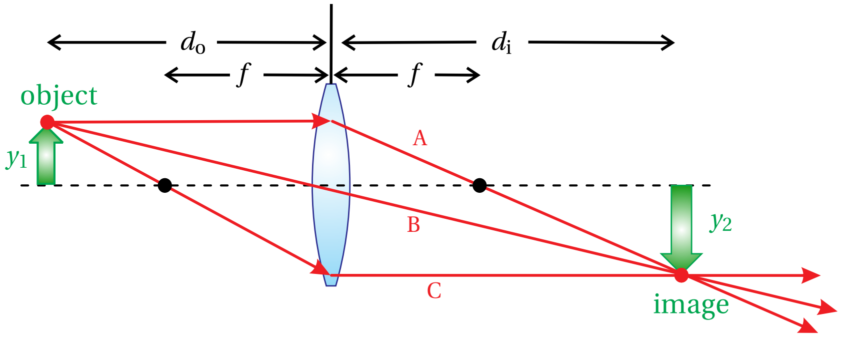


Figure 9.14 Formation of a real image by a thin lens.

Solution: Ray A is parallel to the axis with height y_1 before traversing the lens. Just after the lens, ray A is described by

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ 0 \end{bmatrix} = \begin{bmatrix} y_1 \\ -y_1/f \end{bmatrix}$$

which crosses the axis at the focus $d = f$, since $\begin{bmatrix} 1 & f \\ 0 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ -y_1/f \end{bmatrix} = \begin{bmatrix} 0 \\ -y_1/f \end{bmatrix}$.

Meanwhile, ray B traverses the lens just where it crosses the axis. The lens does nothing to this ray:

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix}\begin{bmatrix} 0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} 0 \\ \theta_1 \end{bmatrix}$$

Ray B is undeflected. For ray B, $\theta_1 = -y_1/d_o$.

Finally, ray C, which crosses the axis a distance $d = f$ before the lens, becomes parallel to the axis after traversing the lens:

$$\begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix}\begin{bmatrix} 1 & f \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} f\theta_1 \\ 0 \end{bmatrix}$$

For ray C, $\theta_1 = (y_2 - y_1)/d_o$.

In the above discussion, we have examined image formation by a thin lens or a curved mirror. Of course, images can also be formed by thick lenses or by more complex composite optical systems (e.g. a system of lenses and spaces). The ABCD matrices for the elements in a composite system are simply multiplied together to obtain an overall ABCD matrix. We can derive the condition for image formation by an arbitrary ABCD matrix in the same way that we did above for a thin lens or curved mirror. As before, consider propagation through a distance d_o from an object to the optical system followed by propagation through a distance d_i to an image. The ABCD matrix for the overall operation is

$$\begin{aligned}\begin{bmatrix} 1 & d_i \\ 0 & 1 \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} 1 & d_o \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} A + d_iC & d_oA + B + d_od_iC + d_iD \\ C & d_oC + D \end{bmatrix} \\ &= \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix}\end{aligned} \tag{9.58}$$

An image occurs according to (9.55) when

general condition for image formation

$$B' = d_o A + B + d_o d_i C + d_i D = 0, \tag{9.59}$$

with magnification

$$M = A' = A + d_i C \tag{9.60}$$

Angular Magnification of a Telescope

A telescope consists of two lenses, the first called the *objective lens* with focal length f_o and the second called the *eye piece* with focal length f_e. The function of a telescope is to map all rays having incident angle θ_1 into corresponding rays all with wider angle θ_2. Importantly, θ_2 should only depend on θ_1 and not on where each ray enters the objective lens (i.e. y_1).

The ABCD matrix of the system is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f_e} & 1 \end{bmatrix} \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{f_o} & 1 \end{bmatrix} = \begin{bmatrix} 1-\frac{L}{f_o} & L \\ -\frac{1}{f_o}-\frac{1}{f_e}+\frac{L}{f_o f_e} & 1-\frac{L}{f_e} \end{bmatrix}$$

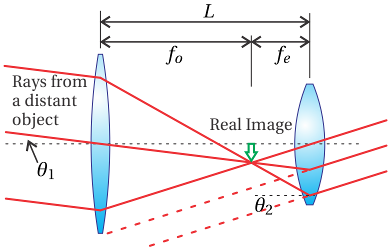


Figure 9.15 Basic telescope consisting of an objective lens and an eye piece.

where L is the separation between the lenses. Since

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} Ay_1 + B\theta_1 \\ Cy_1 + D\theta_1 \end{bmatrix}$$

we need $C = 0$ or $-\frac{1}{f_o} - \frac{1}{f_e} + \frac{L}{f_o f_e} = 0$ to ensure that θ_2 depends only on θ_1. This reduces simply to

$$L = f_o + f_e \tag{9.61}$$

which is the required separation between the objective lens and the eye piece.

The angular magnification is defined to be

$$M_\theta \equiv \frac{\theta_2}{\theta_1} = D = 1 - \frac{L}{f_e} = -\frac{f_o}{f_e} \tag{9.62}$$

The angular magnification is governed by the ratio of the two focal lengths. When looking through a telescope, the apparent angular separation between distant objects is enhanced by this factor. The minus sign indicates that objects in the field of view tend to be inverted.

9.7 Principal Planes for Complex Optical Systems

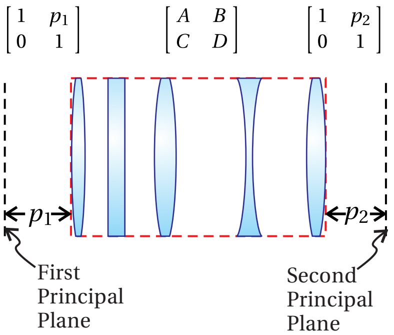
For every ABCD matrix representing an optical system, there exist two *principal planes*,[10] located (in our convention) a distance p_1 before entering the system and a distance p_2 after exiting the system. When the matrices corresponding to these (appropriately chosen) distances are appended to the original ABCD matrix of the system, the overall matrix simplifies to one that looks identical to the matrix for a simple thin lens (9.45).

With knowledge of the positions of the principal planes, one can treat the complicated imaging system in the same way that one treats a simple thin lens. That is, we can simply use the common imaging formulas (9.56) and (9.57). The only difference is that d_o is the distance from the object to the first principal plane and d_i is the distance from the second principal plane to the image. In the case of an actual thin lens, both principal planes are at $p_1 = p_2 = 0$. For a composite system, p_1 and p_2 can be either positive or negative.

We assert that for any optical system,[11] p_1 and p_2 can always be selected such that we can write

$$\begin{aligned}\begin{bmatrix} 1 & p_2 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} 1 & p_1 \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} A + p_2 C & p_1 A + B + p_1 p_2 C + p_2 D \\ C & p_1 C + D \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ -1/f_{\text{eff}} & 1 \end{bmatrix}\end{aligned} \tag{9.63}$$

The final matrix is that of a simple thin lens, and it takes the place of the composite system including the distances to the principal planes.



$\begin{bmatrix} 1 & p_1 \\ 0 & 1 \end{bmatrix}$
$\begin{bmatrix} A & B \\ C & D \end{bmatrix}$
$\begin{bmatrix} 1 & p_2 \\ 0 & 1 \end{bmatrix}$
p_1
p_2
First Principal Plane
Second Principal Plane

Figure 9.16 A multi-element system represented as an ABCD matrix for which principal planes always exist.

Determination of p_1 and p_2 and Justification of (9.63)

Our task is to find the values of p_1 and p_2 that make (9.63) true. We can straightaway make the definition

$$f_{\text{eff}} \equiv -1/C \tag{9.64}$$

We can also solve for p_1 and p_2 by setting the diagonal elements of the matrix to 1. Explicitly, we get

$$p_1 C + D = 1 \quad \Rightarrow \quad p_1 = \frac{1 - D}{C} \tag{9.65}$$

and

$$A + p_2 C = 1 \quad \Rightarrow \quad p_2 = \frac{1 - A}{C} \tag{9.66}$$

It remains to be shown that the upper right element in (9.63) (i.e. $p_1 A + B + p_1 p_2 C + p_2 D$) automatically goes to zero for our choices of p_1 and p_2. This may seem unlikely at first, but watch what happens!

When (9.65) and (9.66) are substituted into the upper right matrix element of (9.63) we get

$$\begin{aligned} p_1 A + B + p_1 p_2 C + p_2 D &= \frac{1-D}{C} A + B + \frac{1-D}{C}\frac{1-A}{C} C + \frac{1-A}{C} D \\ &= \frac{1}{C}\left[1 - AD + BC\right] \\ &= \frac{1}{C}\left(1 - \begin{vmatrix} A & B \\ C & D \end{vmatrix}\right) \end{aligned} \tag{9.67}$$

---

[10] R. Guenther, *Modern Optics*, p. 186 (New York: Wiley, 1990).

[11] The starting and ending refractive index must be the same.

This vanishes (as desired) since the determinant of the ABCD matrix is one, in accordance with (9.52).

9.8 Stability of Laser Cavities

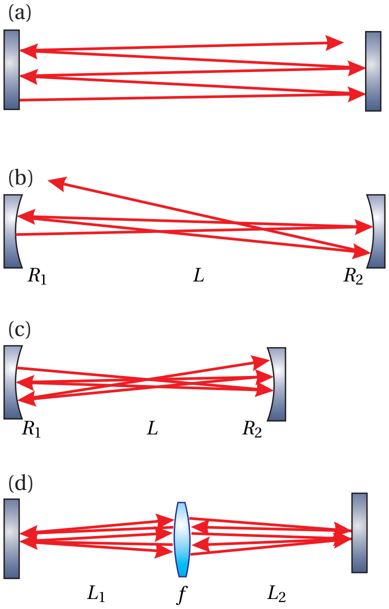


Figure 9.17 (a) A ray bouncing between two parallel flat mirrors. (b) A ray bouncing between two curved mirrors in an unstable configuration. (c) A ray bouncing between two curved mirrors in a stable configuration. (d) Stable cavity utilizing a lens and two flat end mirrors.

The ABCD matrix formulation provides a powerful tool to analyze the *stability* of a *laser cavity*.[12] The basic elements of a laser cavity include an amplifying medium and mirrors to provide feedback. Presumably, at least one of the end mirrors is partially transmitting so that energy is continuously extracted from the cavity. Here, we dispense with the amplifying medium and concentrate our attention on the optics providing the feedback.

As might be expected, the mirrors must be carefully aligned or successive reflections might cause rays to 'walk' continuously away from the optical axis, so that they eventually leave the cavity out the side. If a simple cavity is formed with two flat mirrors that are perfectly aligned parallel to each other, one might suppose that the mirrors would provide ideal feedback. However, all rays except for those that are perfectly aligned to the mirror surface normals would eventually wander out of the side of the cavity as illustrated in Fig. 9.17a. Such a cavity is said to be *unstable*. We would like to do a better job of trapping the light in the cavity.

To improve the situation, a cavity can be constructed with concave end mirrors to help confine the beams within the cavity. Even so, one must choose carefully the curvature of the mirrors and their separation L. If this is not done correctly, the curved mirrors can 'overcompensate' for the tendency of the rays to wander out of the cavity and thus aggravate the problem. Such an unstable scenario is depicted in Fig. 9.17b.

Figure 9.17c depicts a cavity made with curved mirrors where the separation L is chosen appropriately to make the cavity stable. Although a ray, as it makes successive bounces, can strike the end mirrors at a variety of points, the curvature of the mirrors keeps the 'trajectories' contained within a narrow region so that they cannot escape out the sides of the cavity.

There are many ways to make a stable laser cavity. For example, a stable cavity can be made using a lens between two flat end mirrors as shown in Fig. 9.17d. Any combination of lenses (perhaps more than one) and curved mirrors can be used to create stable cavity configurations. *Ring cavities* can also be made to be stable where in no place do the rays retro-reflect from a mirror but circulate through a series of elements like cars going around a racetrack. The ABCD matrix for a round trip in the cavity will be useful for this analysis.

Example 9.9

Find the round-trip ABCD matrix for the cavities shown in Figs. 9.17c and 9.17d.

---

[12] P. W. Milonni and J. H. Eberly, *Lasers*, Sect. 14.3 (New York: Wiley, 1988).

Solution: The round-trip ABCD matrix for the cavity shown in Fig. 9.17c is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -2/R_2 & 1 \end{bmatrix} \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -2/R_1 & 1 \end{bmatrix} \tag{9.68}$$

where we begin the round trip with a reflection from the first mirror.

The round-trip ABCD matrix for the cavity shown in Fig. 9.17d is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 2L_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix} \begin{bmatrix} 1 & 2L_2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix} \tag{9.69}$$

where we begin the round trip with transmission through the lens moving to the right. It is somewhat arbitrary where a round trip begins. Multiplication of the above matrices will be necessary to do problems P9.17 and P9.18.

To determine whether a given configuration of a cavity is stable, we need to know what a ray does after making many round trips in the cavity. To find the effect of propagation through many round trips, we multiply the round-trip ABCD matrix together N times, where N is the number of round trips that we wish to consider. We can then examine what happens to an arbitrary ray after making N round trips in the cavity as follows:

$$\begin{bmatrix} y_{N+1} \\ \theta_{N+1} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^N \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix} \tag{9.70}$$

At this point you might be concerned that taking an ABCD matrix to the N^{th} power can be a lot of work. (It is already a significant work just to compute the ABCD matrix for a single round trip.) In addition, we are interested in letting N be very large, perhaps even infinity. You can relax because we have a neat trick to accomplish this daunting task.

By Sylvester's theorem in appendix 0.3, we have

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^N = \frac{1}{\sin\theta} \begin{bmatrix} A\sin N\theta - \sin(N-1)\theta & B\sin N\theta \\ C\sin N\theta & D\sin N\theta - \sin(N-1)\theta \end{bmatrix} \tag{9.71}$$

where

$$\cos\theta = \frac{1}{2}(A+D) \tag{9.72}$$

This is valid as long as the determinant of the ABCD matrix is one. As noted earlier (see (9.52)), we are in luck! The determinant is one any time a ray begins and stops in the same refractive index, which by definition is guaranteed for any round trip. We therefore can employ Sylvester's theorem for any N that we might choose, including very large integers.

We would like the elements of (9.71) to remain finite as N becomes very large. If this is the case, then we know that a ray remains trapped within the cavity and stays reasonably close to the optical axis. Since N only appears within the argument of a sine function, which is always bounded between -1 and 1 for real arguments, it might seem that the elements of (9.71) always remain finite as N approaches infinity. However, it turns out that θ can become imaginary depending on the outcome of (9.72), in which case the sine becomes a hyperbolic sine, which can 'blow up' as N becomes large. In the end, the condition for cavity stability is that a real θ must exist for (9.72), or in other words we need

$$-1 < \frac{1}{2}(A+D) < 1 \qquad \text{(condition for a stable cavity)} \tag{9.73}$$

It is left as an exercise to apply this condition to (9.68) and (9.69) to find the necessary relationships between the various element curvatures and spacing in order to achieve cavity stability.

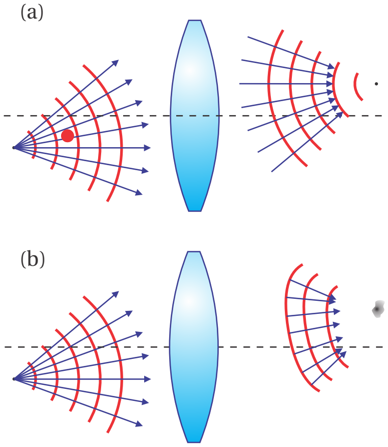


Figure 9.18 (a) Paraxial theory predicts that the light imaged from a point source will converge to a point (i.e. have spherical wavefronts coming to the image point). (b) The image of a point source made by a real lens with aberrations is an extended and blurred patch of light and the converging wavefronts are only quasi-spherical.

Appendix 9.A Aberrations and Ray Tracing

The paraxial approximation places serious limitations on the performance of optical systems (see (9.23) and (9.24)). To stay within the approximation, all rays traveling in the system should travel very close to the optic axis with very shallow angles with respect to the optical axis. To the extent that this is not the case, the collection of rays associated with a single point on an object may not converge to a single point on the associated image. The resulting distortion or "blurring" of the image is known as *aberration*.

Common experience with photographic and video equipment suggests that it is possible to image scenes that have a relatively wide angular extent (many tens of degrees), in apparent serious violation of the paraxial approximation. The paraxial approximation is indeed violated in these devices, so they must be designed using more complicated analysis techniques than those we have learned in this chapter. The most common approach is to use a computationally intensive procedure called *ray tracing* in which $\sin\theta$ and $\tan\theta$ are rendered exactly. The nonlinearity of these functions precludes the possibility of obtaining analytic solutions describing the imaging performance of such optical systems.

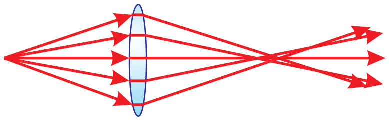


Figure 9.19 Ray tracing through a simple lens.

The typical procedure is to start with a collection of rays from a test point such as shown in Fig. 9.19. Each ray is individually traced through the system using the exact representation of geometric surfaces as well as the exact representation of Snell's law. On close analysis, the rays typically do not converge to a distinct imaging point. Rather, the rays can be 'blurred' out over a range of points where the image is supposed to occur. Depending on the angular distribution of the rays as well as on the elements in the setup, the spread of rays around the image point can be large or small. The engineer who designs the system must determine whether the amount of aberration is acceptable, given the various constraints of the device.

To minimize aberrations below typical tolerance levels, several lenses can be used together. If properly chosen, the lenses (some positive, some negative) separated by specific distances, can result in remarkably low aberration levels over certain ranges of operation for the device. Ray tracing is best done with commercial software designed for this purpose. Such software packages are able to develop and optimize designs for specific applications. A useful feature in many software packages is that the user can specify that the design should employ only standard optical components available from known optics companies. In any case, it is typical to specify that all lenses in the system should have spherical surfaces since these are much less expensive to manufacture. We mention briefly several common classes of aberrations that can occur if a lens system is not properly designed.

Chromatic aberration arises from the fact that the index of refraction for glass varies with the wavelength of light. Since the focal length of a lens depends on the index of refraction (see, for example, Eq. (9.46)), the focal length of a lens varies with the wavelength of light. Chromatic aberration can be compensated for by using a pair of lenses made from two types of glass as shown in Fig. 9.20 (the pair is usually cemented together to form a "doublet" lens). The lens with the shortest focal length is made of the glass whose index has the lesser dependence on wavelength. By properly choosing the prescription of the two lenses, you can exactly compensate for chromatic aberration at two wavelengths and do a good job for a wide range of others. Achromatic doublets can also be designed to minimize spherical aberration (see below), so they are often a good choice when you need a high quality lens.

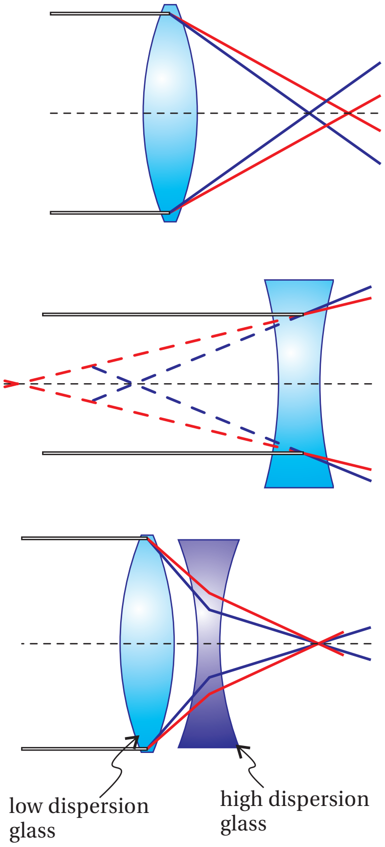


Figure 9.20 Chromatic aberration causes lenses to have different focal lengths for different wavelengths. It can be corrected using an achromatic doublet lens.

Monochromatic aberrations arise from the shape of the lens rather than the variation of *n* with wavelength. Before the advent of computers facilitated the widespread use of ray tracing, these aberrations had to be analyzed analytically. The analytic results derived previously in this chapter were based on a first-order approximation (e.g. $\sin\theta \approx \theta$). One can increase the accuracy of the theory for nonparaxial rays by retaining higher-order terms in the polynomial representation of $\sin\theta$. With higher-order terms included, the wavefronts converging towards an image point are still approximately spherical, but have aberration terms added in (shown conceptually in Fig. 9.18(b)). Without going into detail, there are five aberration terms in the standard second-order analysis, which represent a convenient basis for discussing aberration.

The first aberration term is known as *spherical aberration*. This type of aberration results from the fact that rays traveling through a spherical lens at large radii experience a different focal length than those traveling near the axis. For a converging lens, this causes far-off-radius rays to focus before the near-axis rays as shown in Fig. 9.21. This problem can be helped by orienting lenses so that the face with the least curvature is pointed towards the side where the light rays have the largest angle. This procedure splits the bending of rays more evenly between the front and back surface of the lens. As mentioned above, you can also cement two lenses made from different types of glass together so that spherical aberrations from one lens are corrected by the other.

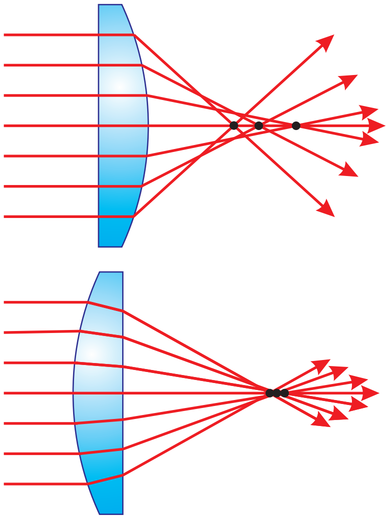


Figure 9.21 Spherical aberration in a plano-convex lens.

The aberration term referred to as *astigmatism* occurs when an off-axis object point is imaged to an off-axis image point. In this case a spherical lens has a different focal length in the horizontal and vertical dimensions. For a focusing lens

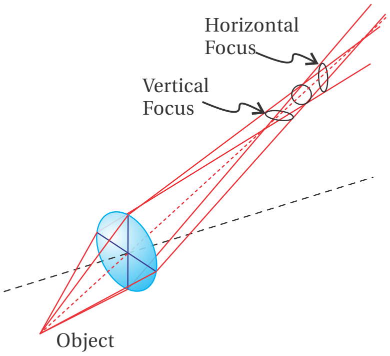


Figure 9.22 Astigmatism causes the horizontal and vertical dimensions to focus at different distances.

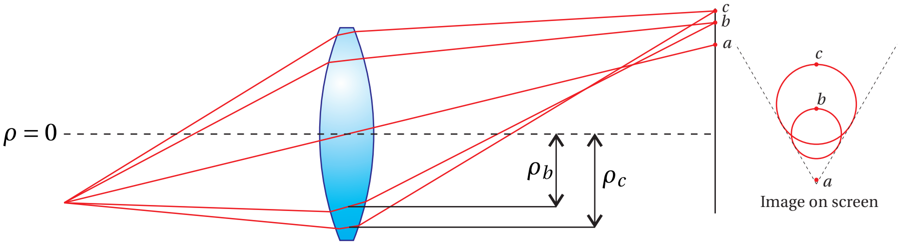


Figure 9.23 Illustration of coma. Rays traveling through the center of the lens are imaged to point a as predicted by paraxial theory. Rays that travel through the lens at radius ρ_b in the plane of the figure are imaged to point b. Rays that travel through the lens at radius ρ_b, but outside the plane of the figure are imaged to other points on the circle (in the image plane) containing point b. Rays at that travel through the lens at other radii on the lens (e.g. ρ_c) also form circles in the image plane with radius proportional to ρ^2 with the center offset from point a a distance proportional to ρ^2. When light from each of these circles combines on the screen it produces an imaged point with a "comet tail."

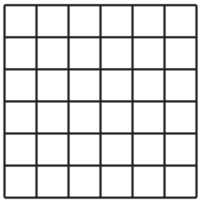
this causes the two dimensions to focus at different distances, producing a vertical 'line' at one image plane and a horizontal 'line' at another (see Fig. 9.22). A lens can also be inherently astigmatic even when viewed on axis if it is football shaped rather than spherical. In this case, the astigmatic aberration can be corrected by inserting a cylindrical lens at the correct orientation (this is a common correction needed in eyeglasses).

A third aberration term is referred to as *coma*. This is observed when off-axis points are imaged and produces a comet shaped tail with its head at the point predicted by paraxial theory. (The term 'coma' refers to the atmosphere of a comet, which is how the aberration got its name.) This aberration is distinct from astigmatism, which is also observed for off-axis points, since coma is observed even when all of the rays are in one plane (see Fig. 9.23). You have probably seen coma if you've ever played with a magnifying glass in the sun—just tilt the lens slightly and you see a comet-like image rather than a point.

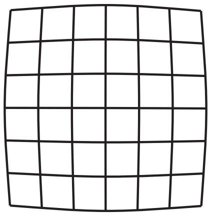
The *curvature of the field* aberration term arises from the fact that spherical lenses image spherical surfaces to another spherical surface, rather than imaging a plane to a plane. This is not so bad for your eyeball, which has a curved screen, but for things like cameras and movie projectors we would like to image to a flat screen. When a flat screen is used and the curvature of the field aberration is present, the image will focus well near the center, but become progressively out of focus as you move to the edge of the screen (i.e. the flat screen is farther from the curved image surface as you move from the center).

The final aberration term is referred to as *distortion*. This aberration occurs when the magnification of a lens depends on the distance from the center of the screen. If magnification decreases as the distance from the center increases, then 'barrel' distortion is observed. When magnification increases with distance, 'pincushion' distortion is observed (see Fig. 9.24).

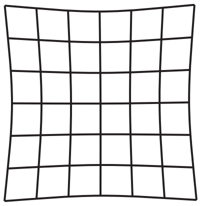
All lenses will exhibit some combination of the aberrations listed above (i.e.



Undistorted



Barrel Distortion



Pincushion Distortion

Figure 9.24 Distortion occurs when magnification is not constant across an extended image.

chromatic aberration plus the five second-order aberration terms). In addition to the five named monochromatic aberrations, there are many other higher order aberrations that also have to be considered. Aberrations can be corrected to a high degree with multiple-element systems (designed using ray-tracing techniques) composed of lenses and irises to eliminate off-axis light. For example, a camera lens with a focal length of 50 mm, one of the simplest lenses in photography, is typically composed of about six individual elements. However, optical systems never completely eliminate all aberration, so designing a system always involves some degree of compromise in choosing which aberrations to minimize and which ones you can live with.

Exercises

Exercises for 9.1 The Eikonal Equation

P9.1 Consider the index described in Example 9.1. The solution given in the example corresponds to rays that asymptotically approach $y = 0$. A more general solution is given by

$$\nabla R = n_0\left(\hat{\mathbf{x}}\sqrt{1+\alpha} \pm \hat{\mathbf{y}}\sqrt{y^2/h^2 - \alpha}\right) \quad \left(1+\alpha > 0 \text{ and } y^2/h^2 - \alpha > 0\right)$$

This corresponds to rays that either hit the ground or return toward the sky without reaching the ground, depending on the sign of α.
(a) Verify that ∇R satisfies the eikonal equation and determine the function $R\left(x, y\right)$.
HINT: $\int d\xi\sqrt{\xi^2 - \alpha} = \frac{\xi}{2}\sqrt{\xi^2-\alpha} - \frac{\alpha}{2}\ln\left(\xi + \sqrt{\xi^2 - \alpha}\right) \quad (\xi - \alpha > 0)$.
(b) Verify that the light path is given by $y = h\sqrt{\alpha}\cosh\left(\frac{x-x_0}{h\sqrt{1+\alpha}}\right)$ when $\alpha > 0$ or is given by $y = h\sqrt{|\alpha|}\sinh\left|\frac{x-x_0}{h\sqrt{1+\alpha}}\right|$ when $\alpha < 0$. Consider only the region $y > 0$ (i.e. above ground). Notice that these solutions can make rays that travel either to the right or to the left.
HINT: $\cosh^2\xi - \sinh^2\xi = 1 \quad \frac{d}{d\xi}\cosh\xi = \sinh\xi \quad \frac{d}{d\xi}\sinh\xi = \cosh\xi$.
(c) Make a sketch of these two solution classes in the case of $\alpha = \pm 1/4$.

P9.2 Prove that under the approximation of very short wavelength, the Poynting vector is directed along $\nabla R(\mathbf{r})$ or $\hat{\mathbf{s}}$.

Solution: (partial)

We have $\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0(\mathbf{r})e^{i(k_{\text{vac}}R(\mathbf{r})-\omega t)}$, and from Faraday's law (1.35) we have

$$\mathbf{B}(\mathbf{r}, t) = -\frac{i}{\omega}\nabla \times \left(\mathbf{E}_0(\mathbf{r})e^{i(k_{\text{vac}}R(\mathbf{r})-\omega t)}\right)$$

Applying the identity $\nabla \times \left(\mathbf{a}\psi\right) = \psi(\nabla \times \mathbf{a}) + \nabla\psi \times \mathbf{a}$ to this equation gives

$$\begin{aligned}
\mathbf{B}(\mathbf{r}, t) &= -\frac{i}{\omega}\left(e^{i(k_{\text{vac}}R(\mathbf{r})-\omega t)}\left[\nabla \times \mathbf{E}_0(\mathbf{r})\right] + i k_{\text{vac}} e^{i(k_{\text{vac}}R(\mathbf{r})-\omega t)}\left[\nabla R(\mathbf{r}) \times \mathbf{E}_0(\mathbf{r})\right]\right) \\
&= -i\frac{\lambda_{\text{vac}}}{2\pi c}e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]}\left[\nabla \times \mathbf{E}_0(\mathbf{r})\right] - \frac{1}{c}e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]}\left[\nabla R(\mathbf{r}) \times \mathbf{E}_0(\mathbf{r})\right]
\end{aligned}$$

The first term vanishes in the limit of very short wavelength, so we simply write

$$\mathbf{B}(\mathbf{r}, t) \to \frac{1}{c}\left[\nabla R(\mathbf{r})\right] \times \mathbf{E}_0(\mathbf{r})\, e^{i[k_{\text{vac}}R(\mathbf{r})-\omega t]}.$$

To obtain the result, we must compute the Poynting vector:

$$\begin{aligned}
\mathbf{S} &= \frac{1}{\mu_0}\text{Re}\{\mathbf{E}(\mathbf{r}, t)\} \times \text{Re}\{\mathbf{B}(\mathbf{r}, t)\} \\
&= \frac{1}{4\mu_0}\left[\mathbf{E}(\mathbf{r}, t) + \mathbf{E}^*(\mathbf{r}, t)\right] \times \left[\mathbf{B}(\mathbf{r}, t) + \mathbf{B}^*(\mathbf{r}, t)\right]
\end{aligned}$$

The BAC-CAB rule (P0.3) will come in handy along with the fact that $\nabla R(\mathbf{r})\cdot\mathbf{E}_0(\mathbf{r})\to 0$. To confirm the latter, we employ Gauss's law (1.33) and the constitutive relation (2.16) as follows:

$$\nabla\cdot\left[\left(1+\chi(\mathbf{r})\right)\mathbf{E}_0(\mathbf{r})e^{i(k_{\text{vac}}R(\mathbf{r})-\omega t)}\right]=0$$

Applying the identity $\nabla\cdot(\mathbf{a}\psi)=\mathbf{a}\cdot\nabla\psi+\psi\nabla\cdot\mathbf{a}$ to this expression yields

$$e^{i(k_{\text{vac}}R(\mathbf{r})-\omega t)}\nabla\cdot\left[\left(1+\chi(\mathbf{r})\right)\mathbf{E}_0(\mathbf{r})\right]+ik_{\text{vac}}e^{i(k_{\text{vac}}R(\mathbf{r})-\omega t)}\left(1+\chi(\mathbf{r})\right)\left[\nabla R(\mathbf{r})\cdot\mathbf{E}_0(\mathbf{r})\right]=0$$

After canceling the common exponential factor, using $k_{\text{vac}}=2\pi/\lambda_{\text{vac}}$, and performing some algebra, we get

$$-i\lambda_{\text{vac}}\frac{\nabla\cdot\left[\left(1+\chi(\mathbf{r})\right)\mathbf{E}_0(\mathbf{r})\right]}{2\pi\left(1+\chi(\mathbf{r})\right)}+\nabla R(\mathbf{r})\cdot\mathbf{E}_0(\mathbf{r})=0$$

In the limit of very short wavelength, this reduces to

$$\nabla R(\mathbf{r})\cdot\mathbf{E}_0(\mathbf{r})\to 0$$

Exercises for 9.2 Fermat's Principle

P9.3 A mirage can be created using a pan of ice water placed above a hotplate with a small air gap,[13] as shown in Fig. 9.25. Suppose a thermocouple measures a uniform temperature gradient from 100°C to 300°C over a distance 4 mm. A narrow laser beam travels $d=16$ cm through the center of the gap where $T=200°\text{C}$. How far is the laser beam deflected laterally (D) after traveling an additional distance $L=40$ m? Assume that the index of refraction follows $n=1+\alpha\frac{P}{T}$, where $\alpha=7.8\times10^{-7}\,\frac{\text{K}}{\text{Pa}}$ and $P=1\text{ atm}=1.013\times10^{5}$ Pa and T is the temperature in Kelvin.

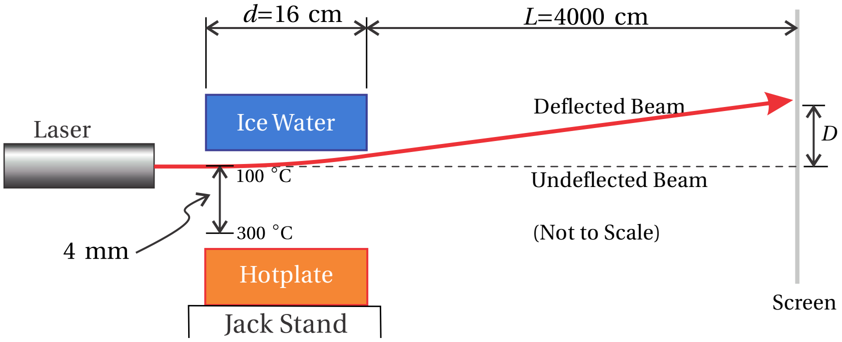


Figure 9.25 Setup to deflect a laser beam between a hotplate and pan of ice water.

HINT: Consider parallel rays within the laser beam, separated by a small lateral displacement Δy. The difference in optical path length ΔOPL while crossing the hotplate compared to Δy matches approximately the lateral displacement D compared to L.

---

[13] L. Richey, B. Stewart, and J. Peatross, "Creating and Analyzing a Mirage," Phys. Teach. **44**, 460-464 (2006).

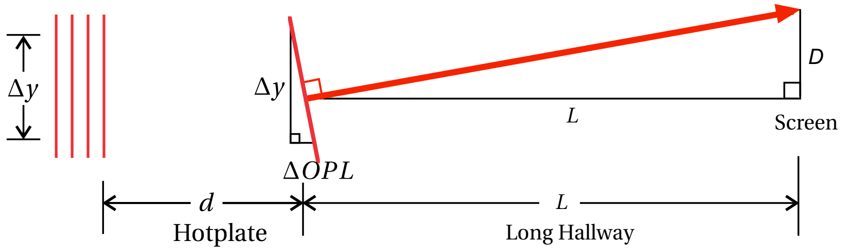


Figure 9.26 Tilting of a laser wavefront in an index gradient.

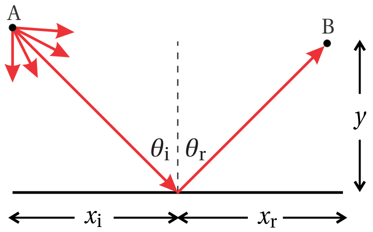


Figure 9.27

P9.4 Use Fermat's Principle to derive the law of reflection (3.6) for a reflective surface.

HINT: Do not consider light that goes directly from A to B; require a single bounce.

P9.5 Show that Fermat's Principle fails to give the correct path for an extraordinary ray entering a uniaxial crystal whose optic axis is perpendicular to the surface.

HINT: With the index given by (5.29), show that Fermat's principle leads to an answer that neither agrees with the direction of the **k**-vector (5.32) nor with the direction of the Poynting vector (5.40).

Exercises for 9.4 Reflection and Refraction at Curved Surfaces

P9.6 Derive the ABCD matrix that takes a ray on a *round trip* through a simple laser cavity consisting of a flat mirror and a concave mirror of radius R separated by a distance L. HINT: Start at the flat mirror. Use the matrix in (9.28) to travel a distance L. Use the matrix in (9.38) to represent reflection from the curved mirror. Then use the matrix in (9.28) to return to the flat mirror. The matrix for reflection from the flat mirror is the identity matrix (i.e. $R_{\text{flat}} \to \infty$).

P9.7 Derive the ABCD matrix for a thick lens made of material n_2 surrounded by a liquid of index n_1. Let the lens have curvatures R_1 and R_2 and thickness d.

Answer:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 + \frac{d}{R_1}\left(\frac{n_1}{n_2} - 1\right) & d\frac{n_1}{n_2} \\ -\left(\frac{n_2}{n_1} - 1\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) + \frac{d}{R_1 R_2}\left(2 - \frac{n_1}{n_2} - \frac{n_2}{n_1}\right) & 1 - \frac{d}{R_2}\left(\frac{n_1}{n_2} - 1\right) \end{bmatrix}$$

When $n_1 = 1$ (i.e. air), the matrix reduces to that in table 9.1.

P9.8 (a) Show that the ABCD matrix for a thick lens given in table 9.1 reduces to that of a thin lens (9.45) when the thickness goes to zero.

(b) Starting from the ABCD matrix of a thick lens in table 9.1, deduce the ABCD matrix for a thick window (thickness d). HINT: A window may be thought of as a thick lens with infinite radii of curvature.

P9.9 Show that the matrix for a thick lens can be derived by sandwiching a window between two thin lenses.

HINT: All relevant formulas appear in table 9.1. Let the thin lenses each have a planar side adjacent to the window. This gives focal lengths $\frac{1}{f_1} = \frac{(n-1)}{R_1}$ and $\frac{1}{f_2} = -\frac{(n-1)}{R_2}$, where R_1 is the radius of the first surface of lens 1, and R_2 is the radius of the second surface of lens 2 (negative if convex).

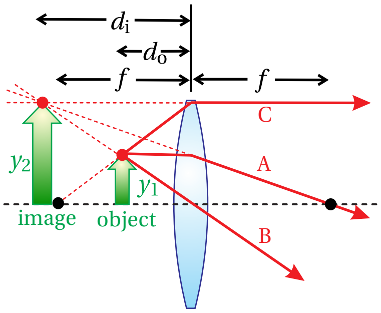


Figure 9.28 Formation of a virtual image by a thin lens.

Exercises for 9.6 Image Formation

P9.10 An object is placed in front of a concave mirror. Find the location of the image d_i and magnification M when $d_o = R$, $d_o = R/2$, $d_o = R/4$, and $d_o = -R/2$ (virtual object). Make a diagram for each situation, depicting rays traveling from a single off-axis point on the object to a corresponding point on the image. You may want to emphasize especially the ray that initially travels parallel to the axis and the ray that initially travels in a direction intersecting the axis at the focal point $R/2$.

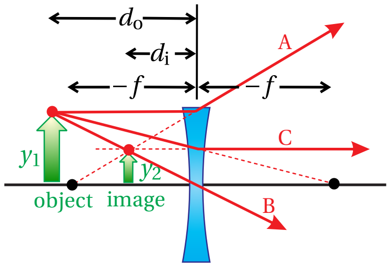


Figure 9.29 Formation of a virtual image by a thin lens with negative focal length.

P9.11 Perform an analysis similar to example 9.8 for the *virtual image* formed by the positive lens in Fig. 9.28. Show that the three final rays all cross at the image.

P9.12 Perform an analysis similar to example 9.8 for the *virtual image* formed by the negative lens in Fig. 9.29. Show that the three final rays all cross at the image.

P9.13 A complicated lens element is represented by an unknown ABCD matrix (see Fig. 9.30). An object placed a distance d_1 before the lens element causes an image to appear a distance d_2 after the unknown element.

Suppose that when $d_1 = \ell$, we find that $d_2 = 2\ell$. Also, suppose that when $d_1 = 2\ell$, we find that $d_2 = 3\ell/2$ with magnification $-1/2$. What is the ABCD matrix for the unknown element?

HINT: Use the conditions for an image (9.59) and (9.60). First find linear expressions for A, B, and C in terms of D. Then put the results into (9.52).

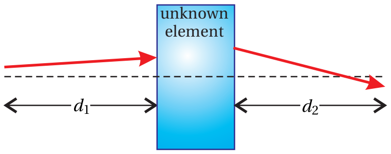


Figure 9.30

Exercises for 9.7 Principal Planes for Complex Optical Systems

P9.14 (a) Consider a thick lens (see Fig. 9.31) with $d = 5$ cm, $R_1 = 5$ cm, $R_2 = -10$ cm, $n = 1.5$. Compute the ABCD matrix of the lens.

(b) Where are the principal planes located and what is the effective focal length f_{eff} for this system?

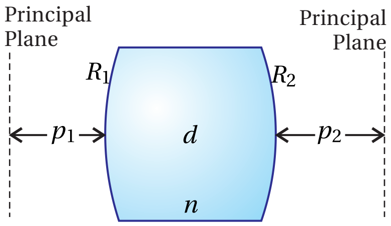


Figure 9.31

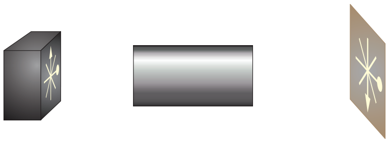


Figure 9.32

L9.15 Deduce the positions of the principal planes and the effective focal length of a compound lens system. Reference the positions of the principal planes to the outside ends of the metal hardware that encloses the lens assembly. (video)

HINT: Obtain three sets of distances to the object and image planes and place the data into (9.59) to create three distinct equations for the unknowns A, B, C, and D. Find A, B, and C in terms of D and place the results into (9.52) to pin down D. The effective focal length and principal planes can then be found through (9.64)–(9.66).

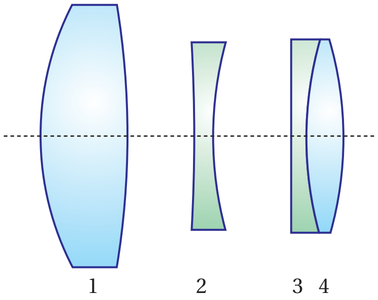


Figure 9.33

P9.16 Use a computer program to calculate the ABCD matrix for the compound system shown in Fig. 9.33, known as the "Tessar lens." The details of this lens are as follows (all distances are in the same units, and *only the magnitude of curvatures are given—you decide the sign*): Convex-convex lens 1 (thickness 0.357, $R_1 = 1.628$, $R_2 = 27.57$, $n = 1.6116$) is separated by 0.189 from concave-concave lens 2 (thickness 0.081, $R_1 = 3.457$, $R_2 = 1.582$, $n = 1.6053$), which is separated by 0.325 from plano-concave lens 3 (thickness 0.217, $R_1 = \infty$, $R_2 = 1.920$, $n = 1.5123$), which is directly followed by convex-convex lens 4 (thickness 0.396, $R_1 = 1.920$, $R_2 = 2.400$, $n = 1.6116$).

Exercises for 9.8 Stability of Laser Cavities

P9.17 (a) Show that the cavity depicted in Fig. 9.17c is stable if

$$0 < \left(1 - \frac{L}{R_1}\right)\left(1 - \frac{L}{R_2}\right) < 1$$

(b) The two concave mirrors have radii $R_1 = 60$ cm and $R_2 = 100$ cm. Over what range of mirror separation L is it possible to form a stable laser cavity?

HINT: There are two different stable ranges with an unstable range between them.

P9.18 Find the stable ranges for $L_1 = L_2 = L$ for the laser cavity depicted in Fig. 9.17d with focal length $f = 50$ cm.

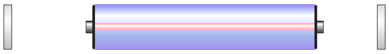


Figure 9.34

L9.19 Experimentally determine the stability range of a HeNe laser with adjustable end mirrors. Check that this agrees reasonably well with theory. Can you think of reasons for any discrepancy? (video)

Chapter 10

Diffraction

In the 1600's, Christiaan Huygens developed a wave description for light. Unfortunately, his ideas were largely overlooked at the time because Sir Isaac Newton promoted a competing theory. Newton proposed that light should be thought of as many tiny bullets, or *corpuscles*, as he called them. Newton's ideas prevailed for more than a century, perhaps because he was right on so many other things, until 1807 when Thomas Young performed his famous two-slit experiment, conclusively demonstrating the wave nature of light. Even then, Young's conclusions were accepted only gradually by others, a notable exception being a young Frenchman named Augustin Fresnel. The two formed a close friendship through correspondence, and it was Fresnel that followed up on Young's conclusions and dedicated his life to a study of light.

Fresnel's skill as a mathematician allowed him to transform physical intuition into powerful and concise ideas. Perhaps Fresnel's greatest accomplishment was the adaptation of *Huygens' principle* of wavelet superposition into a mathematical formula. Ironically, he used Newton's calculus to achieve this. Huygens' principle asserts that a wavefront can be thought of as many wavelets, which propagate and interfere to form new wave fronts. This is illustrated in Fig. 10.1. The phenomenon of diffraction is then understood as the spilling of wavelets around obstructions in the path of light.

After formulating Huygens' principle as a *diffraction integral*, Fresnel made an approximation to his own formula, called the *Fresnel approximation*, for the sake of making the integration easier to perform. As far as approximations go, the Fresnel approximation is surprisingly accurate in describing the light field in the region downstream from an aperture. The diffraction pattern can evolve in complicated ways as the distance from an aperture increases. At distances far downstream from an aperture, the diffraction pattern acquires a final form that no longer evolves, other than to grow in proportion to distance. This *far-field* limit is often of interest, and it turns out that the Fresnel diffraction formula can be simplified further in this case. The far-field limit of the Fresnel diffraction formula is called the Fraunhofer approximation.

From the modern perspective, Fresnel's diffraction formula needs justification

Christiaan Huygens (1629–1695, Dutch) was born in The Hague, Netherlands. His father was friends with the mathematician René Descartes, which probably influenced his upbringing. Huygens studied law and mathematics at the University of Leiden, which preceded a very productive career as a scientist and mathematician. During mid career, Huygens held a position in the French Academy of Sciences in Paris for 15 years, but he spent the majority of his life in The Hague. Huygens was the first to advocate the wave theory of light. He was able to explain birefringence in terms of his wave theory assuming a refractive index that varied with direction. Huygens constructed a telescope with which he discovered Saturn's moon Titan. He also made the first detailed observations of the Orion nebula. Huygens made significant advancements in clock-making technology and wrote a book on probability theory. Huygens was one of the earliest science-fiction writers and speculated that life exists on other planets in his book Cosmotheoros. (Wikipedia) starting from Maxwell’s equations. The diffraction formula is based on *scalar diffraction* theory, which ignores polarization effects. In some situations, ignoring polarization is benign, but in other situations, ignoring polarization effects produces significant errors. These issues as well as the approximations leading to scalar diffraction theory are discussed in section 10.2.

10.1 Huygens’ Principle as Formulated by Fresnel

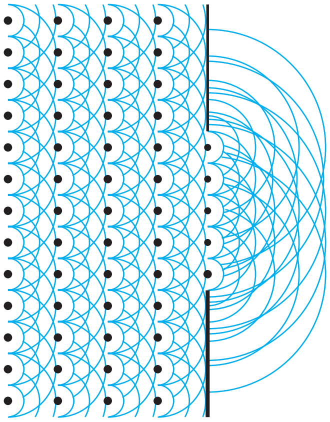
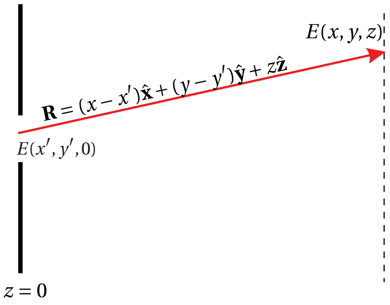


Figure 10.1 Wave fronts depicted as a series of Huygens’ wavelets.

In this section we discuss the calculus of summing up the contributions from the many wavelets originating in an aperture illuminated by a light field. Each point in the aperture is thought of as a source of a *spherical wavelet*.[1] In our modern notation, such a spherical wave can be written as proportional to e^{ikR}/R, where R is the distance from the source. As a spherical wave propagates, its strength falls off in proportion to the distance traveled and the phase is related to the distance propagated, similar to the phase of a plane wave. It should be noted that by choosing k, we consider only a single wavelength of light (i.e. one frequency).

A spherical wave of the form e^{ikR}/R technically does not satisfy Maxwell’s equations (see P10.4). For one thing, it utterly fails near $R = 0$. However, if R is large compared to a wavelength, this spherical wave starts to resemble actual solutions to Maxwell’s equations, as will be examined in the next section. It is within this regime that the diffraction formula derived here is successful.

Consider an aperture or opening in an opaque screen located at the plane $z = 0$. Let the aperture be illuminated with a light field distribution $E(x', y', z = 0)$ within the aperture. Then for a point (x, y, z) lying somewhere after the aperture $(z > 0)$, the net field is given by adding together the contribution of wavelets emitted from each point in the aperture.



$E(x,y,z)$
$\mathbf{R} = (x-x')\hat{\mathbf{x}} + (y-y')\hat{\mathbf{y}} + z\hat{\mathbf{z}}$
$E(x',y',0)$
$z=0$

Figure 10.2

Each spherical wavelet is assigned the strength and phase of the field at the point where it originates. Mathematically, this summation takes the form

$$E(x,y,z) = -\frac{i}{\lambda} \iint\limits_{\text{aperture}} E(x',y',0)\frac{e^{ikR}}{R}\,dx'dy' \tag{10.1}$$

where

$$R = \sqrt{(x-x')^2 + (y-y')^2 + z^2} \tag{10.2}$$

is the radius of each wavelet as it individually intersects the point (x, y, z). We will call (10.1) the Huygens-Fresnel[2] diffraction formula, although Fresnel is credited with this integral formulation. The factor $-i/\lambda$ in front of the integral in (10.1) ensures the right phase and field strength (not to mention correct units). Justification for this factor is given in section 10.3 and in appendix 10.A. To summarize,

---

[1] For simplicity, we use the term ‘spherical wave’ in this book to refer to waves of the type imagined by Huygens (i.e. of the form e^{ikR}/R). There is a different family of waves based on spherical harmonics that are also sometimes referred to as spherical waves. These waves have angular as well as radial dependence, and they *are* solutions to Maxwell’s equations. See J. D. Jackson, *Classical Electrodynamics*, 3rd ed., pp. 429–432 (New York: John Wiley, 1999).

[2] M. Born and E. Wolf, *Principles of Optics*, 7th ed., p. 414 (Cambridge University Press, 1999).

(10.1) tells us how to compute the field downstream given knowledge of the field in an aperture. The field at each point (x', y') in the aperture, which may vary with strength and phase, is treated as the source for a spherical wave. The integral in (10.1) sums the contributions from all of these wavelets.

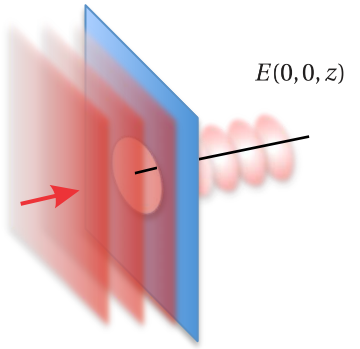


Figure 10.3 Circular aperture illuminated by a plane wave.

Example 10.1

Find the on-axis[3] (i.e. $x, y = 0$) intensity following a circular aperture of *diameter* D illuminated by a uniform plane wave.

Solution: The diffraction integral (10.1) takes the form

$$E(0,0,z) = -\frac{i}{\lambda}\iint\limits_{\text{aperture}} E\left(x', y', 0\right)\frac{e^{ik\sqrt{x'^2+y'^2+z^2}}}{\sqrt{x'^2+y'^2+z^2}}\,dx'dy'$$

The circular hole encourages a change to cylindrical coordinates: $x' = \rho'\cos\phi'$ and $y' = \rho'\sin\phi'$; $dx'dy' \to \rho' d\rho' d\phi'$. In this case, the limits of integration define the geometry of the aperture, and the integration is accomplished as follows:

$$\begin{aligned} E(0,0,z) &= -\frac{iE_0}{\lambda}\int_0^{2\pi} d\phi' \int_0^{D/2} \frac{e^{ik\sqrt{\rho'^2+z^2}}}{\sqrt{\rho'^2+z^2}}\rho'\,d\rho' \\ &= -\frac{iE_0}{\lambda}2\pi \left.\frac{e^{ik\sqrt{\rho'^2+z^2}}}{ik}\right|_0^{D/2} = -E_0\left(e^{ik\sqrt{(D/2)^2+z^2}} - e^{ikz}\right) \end{aligned}$$

The on-axis intensity is then proportional to

$$\begin{aligned} E(0,0,z)\,E^*(0,0,z) &= |E_0|^2\left(e^{ik\sqrt{(D/2)^2+z^2}} - e^{ikz}\right)\left(e^{-ik\sqrt{(D/2)^2+z^2}} - e^{-ikz}\right) \\ &= 2|E_0|^2\left[1-\cos\left(k\sqrt{(D/2)^2+z^2} - kz\right)\right] \end{aligned} \tag{10.3}$$

A graph of this function is shown in Fig. 10.4.

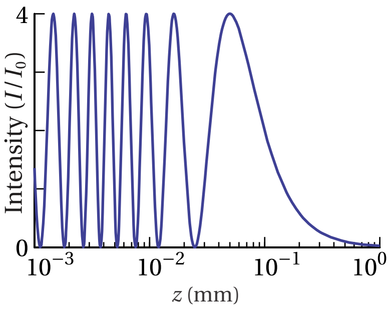


Figure 10.4 Intensity on axis following a circular aperture with $D = 20\lambda$ and wavelength $\lambda = 500$ nm.

When an aperture has a complicated shape, it may be convenient to break up the diffraction integral (10.1) into several pieces. As an example, suppose that we have an aperture consisting of a circular obstruction within a square opening as depicted in Fig. 10.5. Thus, the light transmits through the region between the circle and the square. One can evaluate the overall diffraction pattern by first evaluating the diffraction integral for the entire square (ignoring the circular block) and then subtracting the diffraction integral for a circular opening having the shape of the block. This removes the unwanted part of the previous integration and yields the overall result. When doing this, it is important to add and subtract the integrals (i.e. fields), not their squares (i.e. intensity).

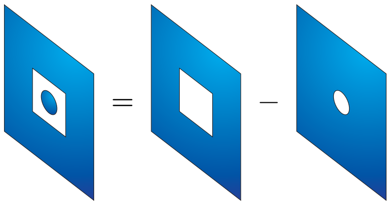


Figure 10.5 Aperture comprised of the region between a circle and a square.

It may be less obvious at first that you can use the above superposition technique to handle diffraction from finite obstructions that interrupt an infinitely wide plane wave. One simply computes the diffraction of the blocked portion of the field as though it came from an opening in a mask. The result is then subtracted from the plane wave (no integration needed for the plane wave), as depicted in Fig. 10.6. This is known as *Babinet's principle*.

---

[3] An analytical solution is not possible off axis.

When Fresnel first presented his diffraction formula to the French Academy of Sciences, a certain judge of scientific papers named Siméon Poisson noticed that Fresnel's formula predicted that there should be light in the center of the geometric shadow behind a circular obstruction. This seemed so absurd to Poisson that he initially disbelieved the theory, until the spot was shortly thereafter experimentally confirmed, much to Poisson's chagrin. Needless to say, Fresnel's paper was then awarded first prize, and this spot appearing behind circular blocks has since been known as *Poisson's spot*.

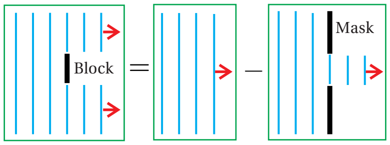


Figure 10.6 Side view of a circular block in a plane wave giving rise to diffraction in the geometric shadow.

Example 10.2

Find the on-axis (i.e. $x, y = 0$) intensity behind a circular block of *diameter* D placed in a uniform plane wave.

Solution: From Example 10.1, the on-axis field behind a circular aperture is $E_0\left(e^{ikz} - e^{ik\sqrt{(D/2)^2+z^2}}\right)$. Babinet's principle says to subtract this result from a plane wave to obtain the field behind the circular block. The situation is depicted in Fig. 10.6. The on-axis field is then

$$E(0,0,z) = E_0 e^{ikz} - E_0\left(e^{ikz} - e^{ik\sqrt{(D/2)^2+z^2}}\right) = E_0 e^{ik\sqrt{(D/2)^2+z^2}}$$

The on axis intensity becomes

$$I(0,0,z) \propto E(0,0,z)\,E^*(0,0,z) = |E_0|^2\, e^{ik\sqrt{(D/2)^2+z^2}} e^{-ik\sqrt{(D/2)^2+z^2}} = |E_0|^2$$

In the exact center of the shadow behind the circular obstruction, the intensity is the same as the illuminating plane wave for all distance z. A spot of light in the center forms right away; no wonder Poisson was astonished!

10.2 Scalar Diffraction Theory

In this section we provide the background motivation for Huygen's principle and Fresnel's formulation of it. Consider a light field with a single frequency ω. The light field can be represented by $\mathbf{E}(\mathbf{r})\, e^{-i\omega t}$, and the time derivative in the wave equation (2.13) can be easily performed. It reduces to

$$\nabla^2\mathbf{E}(\mathbf{r}) + k^2\mathbf{E}(\mathbf{r}) = 0 \tag{10.4}$$

where $k \equiv n\omega/c$ is the magnitude of the usual wave vector (see also (9.2)). Equation (10.4) is called the *Helmholtz equation*. Again, it is merely the wave equation written for the case of a single frequency, where the trivial time dependence has been removed. To obtain the full wave solution, just append the factor $e^{-i\omega t}$ to the solution of (10.4).

At this point we take an *egregious* step: We ignore the vectorial nature of $\mathbf{E}(\mathbf{r})$ and write (10.4) using only the magnitude $E(\mathbf{r})$. When using scalar diffraction theory, we must keep in mind that it is based on this serious step. Under the scalar approximation, the vector Helmholtz equation (10.4) becomes the *scalar Helmholtz equation*:

$$\nabla^2 E(\mathbf{r}) + k^2 E(\mathbf{r}) = 0 \tag{10.5}$$

This equation of course is consistent with (10.4) in the case of a plane wave. However, we are interested in spherical waves of the form $E(r) = E_0 r_0 e^{ikr}/r$. It turns out that such spherical waves are exact solutions to the scalar Helmholtz equation (10.5). The proof is left as an exercise (see P10.3). Nevertheless, spherical waves of this form only approximately satisfy the vector Helmholtz equation (10.4). We can get away with this sleight of hand if the radius r is large compared to a wavelength (i.e. $kr \gg 1$) and if we restrict $\mathbf{r}$ to a narrow range perpendicular to the polarization.

Significance of the Scalar Wave Approximation

The solution of the scalar Helmholtz equation is not completely unassociated with the solution to the vector Helmholtz equation. In fact, if $E_{\text{scalar}}(\mathbf{r})$ obeys the scalar Helmholtz equation (10.5), then

$$\mathbf{E}(\mathbf{r}) = \mathbf{r} \times \nabla E_{\text{scalar}}(\mathbf{r}) \tag{10.6}$$

obeys the vector Helmholtz equation (10.4).

Consider a spherical wave, which is a solution to the scalar Helmholtz equation:

$$E_{\text{scalar}}(\mathbf{r}) = E_0 r_0 e^{ikr}/r \tag{10.7}$$

Remarkably, when this expression is placed into (10.6) the result is zero. Although zero is in fact a solution to the vector Helmholtz equation, it is not very interesting. A more interesting solution to the scalar Helmholtz equation is

$$E_{\text{scalar}}(\mathbf{r}) = r_0 E_0 \left(1 - \frac{i}{kr}\right) \frac{e^{ikr}}{r} \cos\theta \tag{10.8}$$

which is one of an infinite number of unique 'spherical' solutions that exist. Notice that in the limit of large r, this expression looks similar to (10.7), aside from the factor $\cos\theta$. The vector form of this field according to (10.6) is

$$\mathbf{E}(\mathbf{r}) = -\hat{\phi} r_0 E_0 \left(1 - \frac{i}{kr}\right) \frac{e^{ikr}}{r} \sin\theta \tag{10.9}$$

This field looks approximately like the scalar spherical wave solution (10.7) in the limit of large r if the angle is chosen to lie near $\theta \cong \pi/2$ (spherical coordinates). Since our use of the scalar Helmholtz equation is in connection with this spherical wave under these conditions, the results are close to those obtained from the vector Helmholtz equation.

Francois Jean Dominique Arago (1786-1853, French) was born in Catalan France, where his father was the Treasurer of the Mint. As a teenager, Arago was sent to a municipal college in Perpignan where he developed a deep interest in mathematics. In 1803, he entered the École Polytechnique in Paris, where he purportedly was disappointed that he was not presented with new knowledge at a higher rate. He associated with famous French mathematicians Siméon Poisson and Pierre-Simon Laplace. He later worked with Jean-Baptiste Biot to measure the meridian arc to determine the exact length of the meter. This work took him to the Balearic Islands, Spain, where he was imprisoned as a spy, being suspected because of lighting fires atop a mountain as part of his surveying efforts. After a heroic prison escape and a subsequent string of misfortunes, he eventually made it back to France where he took a strong interest in optics and the wave theory of light. Arago and Fresnel established a fruitful collaboration that extended for many years. It was Arago who demonstrated Poisson's spot (sometimes called Arago's spot). Arago also invented the first polarizing filter. In later life, he served a brief stint as the French prime minister. (Wikipedia)

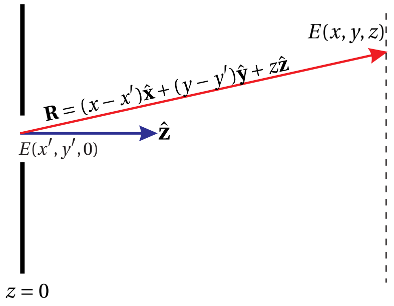

$E(x,y,z)$
$\mathbf{R} = (x - x')\hat{\mathbf{x}} + (y - y')\hat{\mathbf{y}} + z\hat{\mathbf{z}}$
$\hat{\mathbf{z}}$
$E(x',y',0)$
$z = 0$

Figure 10.7

Fresnel developed his diffraction formula (10.1) a half century before Maxwell assembled the equations of electromagnetic theory. In 1887, Gustav Kirchhoff demonstrated that Fresnel's diffraction formula satisfies the scalar Helmholtz equation. In doing this he clearly showed the approximations implicit in the theory, and made a slight revision to the formula:

$$E\left(x,y,z\right) = -\frac{i}{\lambda} \iint\limits_{\text{aperture}} E\left(x',y',0\right) \frac{e^{ikR}}{R} \left[\frac{1+\cos\left(\mathbf{R},\hat{\mathbf{z}}\right)}{2}\right] dx'dy' \tag{10.10}$$

The factor in square brackets, Kirchhoff's revision, is known as the *obliquity factor*. Here, $\cos(\mathbf{R},\hat{\mathbf{z}})$ indicates the cosine of the angle between $\mathbf{R}$ and $\hat{\mathbf{z}}$. Notice that this factor is approximately equal to one when the point (x,y,z) is chosen to be in the forward direction; we usually study diffraction under this circumstance. On the other hand, the obliquity factor equals zero for fields traveling in the reverse direction (i.e. in the $-\hat{\mathbf{z}}$ direction). This fixes a problem with Fresnel's version of the formula (10.1) based on Huygens' wavelets, which suggested that light could as easily diffract in the reverse direction as in the forward direction

In honor of Kirchhoff's work, (10.10) is referred to as the Fresnel-Kirchhoff diffraction formula. The details of Kirchhoff's more rigorous derivation, including how the factor $-i/\lambda$ naturally arises, are given in appendix 10.A. Since the Fresnel-Kirchhoff formula can be understood as a superposition of spherical waves, it is not surprising that it satisfies the scalar Helmholtz equation (10.5).

10.3 Fresnel Approximation

Although the Fresnel-Kirchhoff integral looks innocent enough, it is actually quite difficult to evaluate analytically. Even the Huygens-Fresnel version (10.1) where the obliquity factor $(1+\cos(\mathbf{r},\hat{\mathbf{z}}))/2$ is approximated as one (i.e. far forward direction) is challenging. The integration can be challenging even if we choose a field $E\left(x',y',0\right)$ that is uniform across the aperture (i.e. a constant).

Fresnel introduced an approximation[4] to his diffraction formula that makes the integration somewhat easier to perform. The approximation is analogous to the paraxial approximation made for rays in chapter 9.

Besides letting the obliquity factor be one, Fresnel approximated R by the distance z in the denominator of (10.10) . Then the denominator can be brought out in front of the integral since it no longer depends on x' and y'. This is valid to the extent that we restrict ourselves to small angles:

$$R \cong z \qquad \text{(denominator only; Fresnel approximation)} \tag{10.11}$$

The above approximation is wholly inappropriate in the exponent of (10.10) since small changes in R can result in dramatic variations in the periodic function e^{ikR}.

---

[4] J. W. Goodman, *Introduction to Fourier Optics*, Sect. 4-1 (New York: McGraw-Hill, 1968).

To approximate R in the exponent, we must proceed with caution. To this end we expand (10.2) under the assumption $z^2 \gg (x-x')^2 + (y-y')^2$. Again, this is consistent with the idea of restricting ourselves to relatively small angles. The expansion of (10.2) is written as

$$R = z\sqrt{1 + \frac{(x-x')^2 + (y-y')^2}{z^2}} \cong z\left[1 + \frac{(x-x')^2 + (y-y')^2}{2z^2} + \cdots\right]$$

(exponent; Fresnel approximation) (10.12)

Substitution of (10.11) and (10.12) into the Huygens-Fresnel diffraction formula (10.1) yields

$$E(x,y,z) \cong -\frac{i e^{ikz} e^{i\frac{k}{2z}(x^2+y^2)}}{\lambda z} \iint\limits_{\text{aperture}} E(x',y',0)\, e^{i\frac{k}{2z}(x'^2+y'^2)} e^{-i\frac{k}{z}(xx'+yy')} dx'dy'$$

(Fresnel approximation) (10.13)

This approximation may look a bit messier than before, but in terms of being able to make progress on integration our chances are somewhat improved.

Example 10.3

Compute the Fresnel diffraction field following a rectangular aperture (dimensions Δx by Δy) illuminated by a uniform plane wave.

Solution: According to (10.13), the field downstream is

$$E(x,y,z) = -iE_0 \frac{e^{ikz}}{\lambda z} e^{i\frac{k}{2z}(x^2+y^2)} \int_{-\Delta x/2}^{\Delta x/2} dx' e^{i\frac{k}{2z}x'^2} e^{-i\frac{kx}{z}x'} \int_{-\Delta y/2}^{\Delta y/2} dy' e^{i\frac{k}{2z}y'^2} e^{-i\frac{ky}{z}y'}$$

Unfortunately, the integration in the preceding example must be performed numerically. This is often the case for diffraction integrals in the Fresnel approximation, but at least numerical fast Fourier transforms can aid in the process. Figure 10.8 shows the result of integration for a rectangular aperture with a height twice its width.

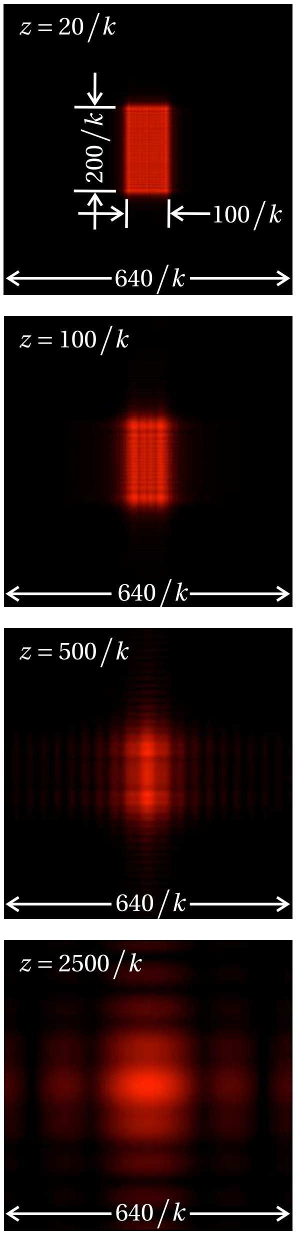


Figure 10.8 Field amplitude following a rectangular aperture computed in the Fresnel approximation.

Paraxial Wave Equation

If we assume that the light coming through the aperture is highly directional, such that it propagates mainly in the z-direction, we are motivated to write the field as $E(x,y,z) = \tilde{E}(x,y,z)e^{ikz}$. Upon substitution of this into the scalar Helmholtz equation (10.5), we arrive at

$$\frac{\partial^2 \tilde{E}}{\partial x^2} + \frac{\partial^2 \tilde{E}}{\partial y^2} + 2ik\frac{\partial \tilde{E}}{\partial z} + \frac{\partial^2 \tilde{E}}{\partial z^2} = 0 \qquad (10.14)$$

At this point we make the paraxial wave approximation,[5] which is $|2k\frac{\partial \tilde{E}}{\partial z}| \gg |\frac{\partial^2 \tilde{E}}{\partial z^2}|$. That is, we assume that the amplitude of the field varies slowly in the z-direction such that the wave looks much like a plane wave. We permit the amplitude to change as the wave propagates in the z-direction as long as it does so on a scale much longer than a wavelength. This leads to the *paraxial wave equation*:

---

[5] P. W. Milonni and J. H. Eberly, *Laser*, Sect. 14.4 (New York: Wiley, 1988).

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+2ik\frac{\partial}{\partial z}\right)\tilde{E}(x,y,z)\cong 0 \qquad \text{(paraxial wave equation)} \quad (10.15)$$

It turns out that the Fresnel approximation (10.13) is an *exact* solution to the paraxial wave equation (see P10.5). That is, (10.15) is satisfied by

$$\tilde{E}(x,y,z)\cong -\frac{i}{\lambda z}\iint\limits_{-\infty}^{\infty}\tilde{E}(x',y',0)\,e^{i\frac{k}{2z}\left[(x-x')^2+(y-y')^2\right]}dx'dy' \qquad (10.16)$$

When the factor e^{ikz} is appended, this field is identical to (10.13).

Joseph von Fraunhofer (1787–1826, German) was born in Straubing, Bavaria. He was orphaned at age 11, whereupon he was apprenticed to a glassmaker. The workshop collapsed, trapping him in the rubble. The Prince of Bavaria directed the rescue efforts and thereafter took an interest in Fraunhofer's education. The prince required the glassmaker to allow young Joseph time to study, and he naturally took an interest in optics. Fraunhofer later worked at the Optical Institute at Benediktbeuern, where he learned techniques for making the finest optical glass in his day. Fraunhofer developed numerous glass recipes and was expert at creating optical devices. Fraunhofer was the inventor of the spectroscope, making it possible to do quantitative spectroscopy. Using his spectroscope, Fraunhofer was the first to observe and document hundreds of absorption lines in the sun's spectrum. He also noticed that these varied for different stars, thus establishing the field of stellar spectroscopy. He was also the inventor of the diffraction grating. In 1822, he was granted an honorary doctorate from the University of Erlangen. Fraunhofer passed away at age 39, perhaps due to heavy-metal poisoning from glass blowing. (Wikipedia)

10.4 Fraunhofer Approximation

An additional approximation to the diffraction integral was made famous by Joseph von Fraunhofer. The *Fraunhofer approximation* is the limiting case of the Fresnel approximation when the field is observed at a distance far after the aperture (called the *far field*).[6] A diffraction pattern continuously evolves along the z-direction, as described by the Fresnel approximation. Eventually it evolves into a final diffraction pattern that maintains itself as it continues to propagate (although it increases its size in proportion to distance). It is this far-away diffraction pattern that is obtained from the Fraunhofer approximation. Since the Fresnel approximation requires the angles to be small (i.e. the paraxial approximation), so does the Fraunhofer approximation.

To obtain the diffraction pattern at a distance very far from the aperture, we make the following approximation:[7]

$$e^{i\frac{k}{2z}\left(x'^2+y'^2\right)}\cong 1 \qquad \text{(far field)} \quad (10.17)$$

The validity of this approximation depends on a comparison of the size of the aperture to the distance z where the diffraction pattern is observed. We need

$$z\gg\frac{k}{2}\left(\text{aperture radius}\right)^2 \qquad \text{(condition for far field)} \quad (10.18)$$

By removing the factor (10.17) from (10.13), we obtain the Fraunhofer diffraction formula:

(Fraunhofer approximation)

$$E\left(x,y,z\right)\cong -\frac{ie^{ikz}e^{i\frac{k}{2z}\left(x^2+y^2\right)}}{\lambda z}\iint\limits_{\text{aperture}}E\left(x',y',0\right)e^{-i\frac{k}{z}\left(xx'+yy'\right)}dx'dy' \qquad (10.19)$$

---

[6] Since the Fraunhofer approximation is easier to use, many textbooks present it before the Fresnel approximation.

[7] J. W. Goodman, *Introduction to Fourier Optics*, p. 61 (New York: McGraw-Hill, 1968).

Obviously, the removal of $e^{i\frac{k}{2z}\left(x'^2+y'^2\right)}$ from the integrand improves our chances of being able to perform the integration analytically. In fact the integral can be interpreted as a two-dimensional (inverse) Fourier transform on the aperture field $E\left(x',y',0\right)$, where kx/z and ky/z can be thought of as 'spatial frequencies'.

Once we are in the regime where the Fraunhofer approximation is valid, a change in z is not very interesting since it appears within the integral only in the combination x/z or y/z. At a larger distance z, the same diffraction pattern is obtained with a proportionately larger values of x or y. The Fraunhofer diffraction pattern thus preserves itself indefinitely as the field propagates. It grows in size as the distance z increases, but the *angular* size defined by x/z or y/z remains the same.

Example 10.4

Compute the Fraunhofer diffraction pattern following a rectangular aperture (dimensions Δx by Δy) illuminated by a uniform plane wave.

Solution: According to (10.19), the field downstream is

$$E\left(x,y,z\right)=-iE_0\frac{e^{ikz}}{\lambda z}e^{i\frac{k}{2z}\left(x^2+y^2\right)}\int\limits_{-\Delta x/2}^{\Delta x/2}dx'e^{-i\frac{kx}{z}x'}\int\limits_{-\Delta y/2}^{\Delta y/2}dy'e^{-i\frac{ky}{z}y'}$$

It is left as an exercise (see P10.6) to perform the integration and compute the intensity. The result turns out to be

$$I\left(x,y,z\right)=I_0\frac{\Delta x^2\Delta y^2}{\lambda^2z^2}\text{sinc}^2\left(\frac{\pi\Delta x}{\lambda z}x\right)\text{sinc}^2\left(\frac{\pi\Delta y}{\lambda z}y\right) \tag{10.20}$$

where $\text{sinc}\left(\xi\right)\equiv\sin\xi/\xi$. Note that $\lim\limits_{\xi\to0}\text{sinc}\left(\xi\right)=1$.

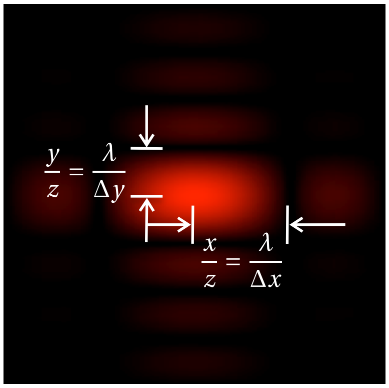


Figure 10.9 Fraunhofer diffraction pattern (field amplitude) generated by a uniformly illuminated rectangular aperture with a height twice the width.

10.5 Diffraction with Cylindrical Symmetry

Sometimes the field transmitted by an aperture is cylindrically symmetric. In this case, the field at the aperture can be written as

$$E(x',y',z=0)=E(\rho',z=0) \tag{10.21}$$

where $\rho\equiv\sqrt{x^2+y^2}$. Under cylindrical symmetry, the two-dimensional integration over x' and y' in (10.13) or (10.19) can be reduced to a single-dimensional integral over a cylindrical coordinate ρ'. With the coordinate transformation

$$x\equiv\rho\cos\phi \qquad y\equiv\rho\sin\phi \qquad x'\equiv\rho'\cos\phi' \qquad y'\equiv\rho'\sin\phi' \tag{10.22}$$

the Fresnel diffraction integral (10.13) becomes

$$E\left(\rho,z\right)=-\frac{ie^{ikz}e^{i\frac{k\rho^2}{2z}}}{\lambda z}\int\limits_0^{2\pi}d\phi'\int\limits_{\text{aperture}}\rho'd\rho'E\left(\rho',0\right)e^{i\frac{k\rho'^2}{2z}}e^{-i\frac{k}{z}\left(\rho\rho'\cos\phi\cos\phi'+\rho\rho'\sin\phi\sin\phi'\right)} \tag{10.23}$$

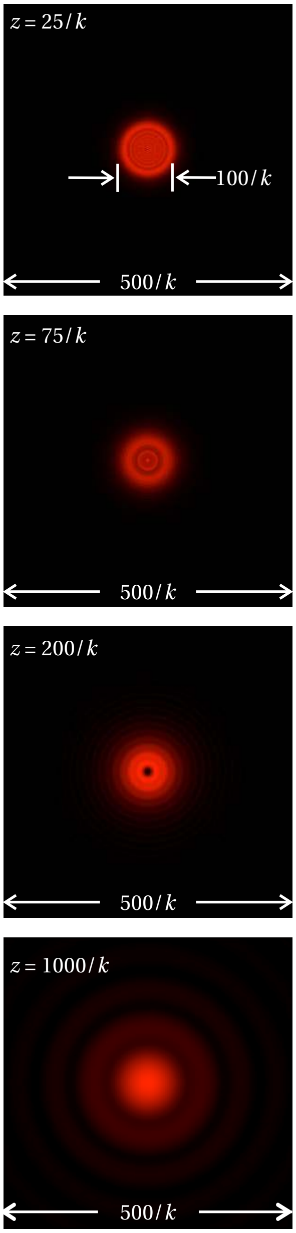


Figure 10.10 Field amplitude following a circular aperture computed in the Fresnel approximation.

Notice that in the exponent of (10.23) we can write

$$\rho'\rho\left(\cos\phi'\cos\phi+\sin\phi'\sin\phi\right)=\rho'\rho\cos\left(\phi'-\phi\right) \tag{10.24}$$

With this simplification, the diffraction formula (10.23) can be written as

$$E\left(\rho,z\right)=-\frac{ie^{ikz}e^{i\frac{k\rho^2}{2z}}}{\lambda z}\int\limits_{\text{aperture}}\rho'd\rho'E\left(\rho',0\right)e^{i\frac{k\rho'^2}{2z}}\int\limits_0^{2\pi}d\phi'e^{-i\frac{k\rho\rho'}{z}\cos(\phi-\phi')} \tag{10.25}$$

We are able to perform the integration over ϕ' with the help of the formula (0.57):

$$\int\limits_0^{2\pi}e^{-i\frac{k\rho\rho'}{z}\cos(\phi-\phi')}d\phi'=2\pi J_0\left(\frac{k\rho\rho'}{z}\right) \tag{10.26}$$

J_0 is called the zero-order *Bessel function.* Equation (10.25) then reduces to

$$E\left(\rho,z\right)=-\frac{2\pi ie^{ikz}e^{i\frac{k\rho^2}{2z}}}{\lambda z}\int\limits_{\text{aperture}}\rho'd\rho'E\left(\rho',0\right)e^{i\frac{k\rho'^2}{2z}}J_0\left(\frac{k\rho\rho'}{z}\right)$$

(Fresnel approximation with cylindrical symmetry) (10.27)

The integral in (10.27) is called a *Hankel transform* on $E\left(\rho',0\right)e^{i\frac{k\rho'^2}{2z}}$.

In the case of the Fraunhofer approximation, the diffraction integral becomes a Hankel transform on just the field $E\left(\rho',z=0\right)$ since $\exp\left(i\frac{k\rho'^2}{2z}\right)$ goes to one. Under cylindrical symmetry, the Fraunhofer approximation is

$$E\left(\rho,z\right)=-\frac{2\pi ie^{ikz}e^{i\frac{k\rho^2}{2z}}}{\lambda z}\int\limits_{\text{aperture}}\rho'd\rho'E\left(\rho',0\right)J_0\left(\frac{k\rho\rho'}{z}\right)$$

(Fraunhofer approximation with cylindrical symmetry) (10.28)

Just as fast Fourier transform algorithms aid in the numerical evaluation of diffraction integrals in Cartesian coordinates, fast Hankel transforms also exist and can be used with cylindrically symmetric diffraction integrals.

Example 10.5

Compute the Fresnel and Fraunhofer diffraction patterns following a circular aperture (diameter D) illuminated by a uniform plane wave.

Solution: According to (10.27), the field downstream is

$$E\left(\rho,z\right)=-iE_0\frac{2\pi e^{ikz}e^{i\frac{k\rho^2}{2z}}}{\lambda z}\int\limits_0^{D/2}\rho'd\rho'e^{i\frac{k\rho'^2}{2z}}J_0\left(\frac{k\rho\rho'}{z}\right)$$

Unfortunately, this Fresnel integral must be performed numerically. The result of the calculation for a uniform field illuminating a circular aperture is shown in Fig. 10.10.

On the other hand, the field in the Fraunhofer limit (10.28) is

$$E\left(\rho,z\right) = -iE_0\frac{2\pi e^{ikz}e^{i\frac{k\rho^2}{2z}}}{\lambda z}\int\limits_0^{D/2}\rho' d\rho' J_0\left(\frac{k\rho\rho'}{z}\right)$$

which *can* be integrated analytically (with the aid of (0.58)). It is left as an exercise to perform the integration and to show that the intensity of the Fraunhofer pattern is

$$I\left(\rho,z\right) = I_0\left(\frac{\pi D^2}{4\lambda z}\right)^2\left[2\frac{J_1\left(kD\rho/2z\right)}{\left(kD\rho/2z\right)}\right]^2 \tag{10.29}$$

The function $\frac{2J_1(\xi)}{\xi}$, which we will call the *jinc* function,[8] looks similar to the sinc function (see Example 10.4) except that its first zero is at $\xi = 1.22\pi$ rather than at π. Note that $\lim\limits_{\xi\to 0}\frac{2J_1(\xi)}{\xi} = 1$.

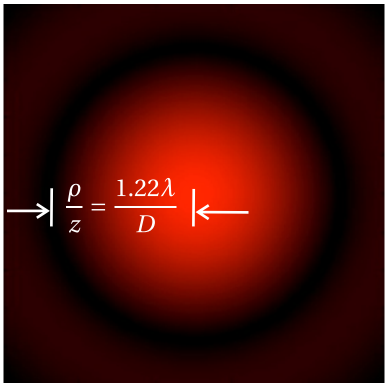


Figure 10.11 Fraunhofer diffraction pattern (field amplitude) generated for a uniformly illuminated circular aperture.

Appendix 10.A Fresnel-Kirchhoff Diffraction Formula

To begin the derivation of the Fresnel-Kirchhoff diffraction formula,[9] we employ Green's theorem (proven in appendix 10.B):

$$\oint\limits_S\left[U\frac{\partial V}{\partial n} - V\frac{\partial U}{\partial n}\right]da = \int\limits_V\left[U\nabla^2 V - V\nabla^2 U\right]dv \tag{10.30}$$

The notation $\partial/\partial n$ implies a derivative in the direction normal to the surface. We choose the following functions:

$$\begin{aligned} V &\equiv e^{ikr}/r \\ U &\equiv E\left(\mathbf{r}\right) \end{aligned} \tag{10.31}$$

where $E\left(\mathbf{r}\right)$ is assumed to satisfy the scalar Helmholtz equation, (10.5). When these functions are used in Green's theorem (10.30), we obtain

$$\oint\limits_S\left[E\frac{\partial}{\partial n}\frac{e^{ikr}}{r} - \frac{e^{ikr}}{r}\frac{\partial E}{\partial n}\right]da = \int\limits_V\left[E\nabla^2\frac{e^{ikr}}{r} - \frac{e^{ikr}}{r}\nabla^2 E\right]dv \tag{10.32}$$

The right-hand side of this equation vanishes[10] since we have

$$E\nabla^2\frac{e^{ikr}}{r} - \frac{e^{ikr}}{r}\nabla^2 E = -k^2 E\frac{e^{ikr}}{r} + \frac{e^{ikr}}{r}k^2 E = 0 \tag{10.33}$$

---

[8] Most authors define the jinc without the factor of 2, which gives the inconvenient normalization $\lim\limits_{\xi\to 0}\text{jinc}(\xi) = 1/2$.

[9] See J. W. Goodman, *Introduction to Fourier Optics*, Sect. 3-3 (New York: McGraw-Hill, 1968).

[10] We exclude the point $r = 0$; see P0.4 and P0.5.

where we have taken advantage of the fact that $E(\mathbf{r})$ and e^{ikr}/r both satisfy (10.5). This is exactly the reason for our judicious choices of the functions V and U since with them we were able to make half of (10.30) disappear. We are left with

$$\oint_S \left[E\frac{\partial}{\partial n}\frac{e^{ikr}}{r} - \frac{e^{ikr}}{r}\frac{\partial E}{\partial n}\right] da = 0 \tag{10.34}$$

Now consider a volume between a small sphere of radius ϵ at the origin and an outer surface of arbitrary shape. The total surface that encloses the volume is comprised of two parts (i.e. $S = S_1 + S_2$ as depicted in Fig. 10.12).

When we apply (10.34) to the surface in Fig. 10.12, we have

$$\oint_{S_2} \left[E\frac{\partial}{\partial n}\frac{e^{ikr}}{r} - \frac{e^{ikr}}{r}\frac{\partial E}{\partial n}\right] da = -\oint_{S_1} \left[E\frac{\partial}{\partial n}\frac{e^{ikr}}{r} - \frac{e^{ikr}}{r}\frac{\partial E}{\partial n}\right] da \tag{10.35}$$

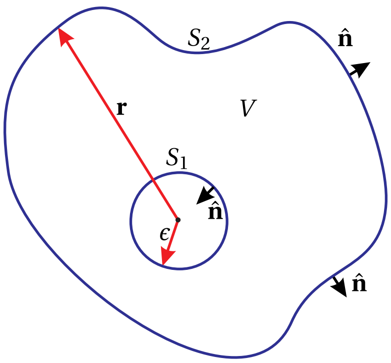


Figure 10.12 A two-part surface enclosing volume V.

This geometry with multiple surfaces is motivated by the hope of finding the field at the origin (inside the little sphere) from knowledge of the field on the outside surface. To this end, we assume that ϵ is so small that $E(\mathbf{r})$ is approximately the same everywhere on the surface S_1. Then the integral over S_1 becomes

$$\oint_{S_1} \left[E\frac{\partial}{\partial n}\frac{e^{ikr}}{r} - \frac{e^{ikr}}{r}\frac{\partial E}{\partial n}\right] da = \lim_{r=\epsilon\to 0}\int_0^{2\pi} d\phi \int_0^{\pi} \left[E\left(\frac{\partial}{\partial r}\frac{e^{ikr}}{r}\right)\frac{\partial r}{\partial n} - \frac{e^{ikr}}{r}\left(\frac{\partial E}{\partial r}\right)\frac{\partial r}{\partial n}\right] r^2 \sin\theta\, d\theta \tag{10.36}$$

where we have used spherical coordinates. Notice that we have employed the chain rule to execute the normal derivative $\partial/\partial n$. Since r always points opposite to the direction of the surface normal $\hat{\mathbf{n}}$, the normal derivative $\partial r/\partial n$ is always equal to -1.[11] We can perform the angular integration in (10.36) as well as take the limit $\epsilon \to 0$:

$$\begin{aligned}\lim_{\epsilon\to 0}\oint_{S_1} \left[E\frac{\partial}{\partial n}\frac{e^{ikr}}{r} - \frac{e^{ikr}}{r}\frac{\partial E}{\partial n}\right] da &= -4\pi \lim_{\epsilon\to 0}\left[r^2\left(-\frac{e^{ikr}}{r^2} + ik\frac{e^{ikr}}{r}\right)E - r^2\frac{e^{ikr}}{r}\left(\frac{\partial E}{\partial r}\right)\right]_{r=\epsilon} \\ &= -4\pi \lim_{\epsilon\to 0}\left[\left(-e^{ik\epsilon} + ik\epsilon e^{ik\epsilon}\right)E - e^{ik\epsilon}\epsilon\left(\frac{\partial E}{\partial r}\right)_{r=\epsilon}\right] \\ &= 4\pi E(0)\end{aligned} \tag{10.37}$$

With the aid of (10.37), Green's theorem applied to our specific geometry reduces to

$$E(0) = \frac{1}{4\pi}\oint_{S_2}\left[\frac{e^{ikr}}{r}\frac{\partial E}{\partial n} - E\frac{\partial}{\partial n}\frac{e^{ikr}}{r}\right] da \tag{10.38}$$

---

[11] From the definition of the normal derivative we have $\partial r/\partial n \equiv \nabla r \cdot \hat{\mathbf{n}} = -\hat{\mathbf{n}}\cdot\hat{\mathbf{n}} = -1$.

If we know E everywhere on the outer surface S_2, this equation allows us to predict the field $E(0)$ at the origin.

Now let us choose a specific surface S_2. Consider an infinite mask with a finite aperture connected to a hemisphere of infinite radius $R \to \infty$. In the end, we will suppose that light enters through the mask and propagates to our origin (among other points). In our present coordinate system, the vectors $\mathbf{r}$ and $\hat{\mathbf{n}}$ point opposite to the incoming light.

We must evaluate (10.38) on the surface depicted in the figure. For the portion of S_2 that is on the hemisphere, the integrand tends to zero as R becomes large. To argue this, it is necessary to recognize the fact that at large distances the field takes on a form proportional to e^{ikr}/r so that the two terms in the integrand cancel. On the mask, we assume, as did Kirchhoff, that both $\partial E/\partial n$ and E are zero.[12] Thus, we are left with only the integration over the open aperture:

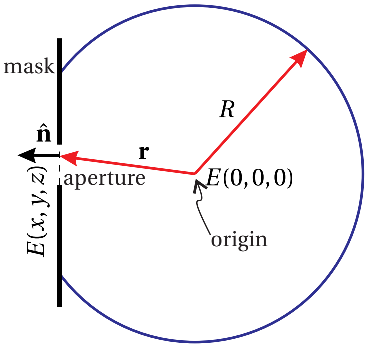


Figure 10.13 Surface S_2 depicted as a mask and a large hemisphere.

$$E(0) = \frac{1}{4\pi} \iint\limits_{\text{aperture}} \left[\frac{e^{ikr}}{r} \frac{\partial E}{\partial n} - E \frac{\partial}{\partial n} \frac{e^{ikr}}{r} \right] da \tag{10.39}$$

We have essentially arrived at the result that we are seeking. The field coming through the aperture is integrated to find the field at the origin, which is located beyond the aperture. Let us manipulate the formula a little further. The second term in the integral of (10.39) can be rewritten as follows:

$$\frac{\partial}{\partial n} \frac{e^{ikr}}{r} = \left(\frac{\partial}{\partial r} \frac{e^{ikr}}{r} \right) \frac{\partial r}{\partial n} = \left(\frac{ik}{r} - \frac{1}{r^2} \right) e^{ikr} \cos(\mathbf{r}, \hat{\mathbf{n}}) \underset{r \gg \lambda}{\rightarrow} \frac{ike^{ikr}}{r} \cos(\mathbf{r}, \hat{\mathbf{n}}) \tag{10.40}$$

where $\partial r/\partial n = \cos(\mathbf{r}, \hat{\mathbf{n}})$ indicates the cosine of the angle between $\mathbf{r}$ and $\hat{\mathbf{n}}$. We have also assumed that the distance r is much larger than a wavelength in order to drop a term. Next, we assume that the field illuminating the aperture can be written as $E \cong \tilde{E}(x, y) e^{ikz}$. This represents a plane-wave field traveling through the aperture from left to right. Then, we have

$$\frac{\partial E}{\partial n} = \frac{\partial E}{\partial z} \frac{\partial z}{\partial n} = ik\tilde{E}(x, y) e^{ikz} (-1) = -ikE \tag{10.41}$$

Substituting (10.40) and (10.41) into (10.39) yields

$$E(0) = -\frac{i}{\lambda} \iint\limits_{\text{aperture}} E \frac{e^{ikr}}{r} \left[\frac{1 + \cos(\mathbf{r}, \hat{\mathbf{n}})}{2} \right] da \tag{10.42}$$

Finally, we wish to rearrange our coordinate system to that depicted in Fig. 10.2. In our derivation, it was less cumbersome to place the origin at a point of interest after the aperture. Now that we have completed our mathematics, we can switch around the coordinate system and place the origin in the plane of the aperture as in Fig. 10.2:

---

[12] Later Sommerfeld noticed that these two assumptions actually contradict each other, and he revised Kirchhoff's work to be more accurate. In practice this revision makes only a tiny difference as light spills onto the back of the aperture, over a length scale of a wavelength. We will ignore this effect and go with Kirchhoff's (slightly flawed) assumption. For further discussion see J. W. Goodman, *Introduction to Fourier Optics*, Sect. 3-4 (New York: McGraw-Hill, 1968).

$$E\left(x,y,z\right)=-\frac{i}{\lambda}\iint\limits_{\text{aperture}} E\left(x',y',0\right)\frac{e^{ikR}}{R}\left[\frac{1+\cos\left(\mathbf{r},\hat{\mathbf{z}}\right)}{2}\right]dx'dy' \tag{10.43}$$

where

$$R=\sqrt{(x-x')^2+\left(y-y'\right)^2+z^2} \tag{10.44}$$

which brings us to the Fresnel-Kirchhoff diffraction formula (10.10).

Appendix 10.B Green's Theorem

To derive Green's theorem, we begin with the divergence theorem (see (0.11)):

$$\oint\limits_S \mathbf{f}\cdot\hat{\mathbf{n}}\,da=\int\limits_V \nabla\cdot\mathbf{f}\,dv \tag{10.45}$$

The unit vector $\hat{\mathbf{n}}$ always points normal to the surface of volume V over which the integral is taken. Let the vector function $\mathbf{f}$ be $U\nabla V$, where U and V are both analytical functions of the position coordinate $\mathbf{r}$. Then (10.45) becomes

$$\oint\limits_S (U\nabla V)\cdot\hat{\mathbf{n}}\,da=\int\limits_V \nabla\cdot(U\nabla V)\;dv \tag{10.46}$$

We recognize $\nabla V\cdot\hat{\mathbf{n}}$ as the directional derivative of V, directed along the surface normal $\hat{\mathbf{n}}$. This is often represented in shorthand notation as

$$\nabla V\cdot\hat{\mathbf{n}}\equiv\frac{\partial V}{\partial n} \tag{10.47}$$

The integrand on the right-hand side of (10.46) can be expanded with the product rule:

$$\nabla\cdot(U\nabla V)=\nabla U\cdot\nabla V+U\nabla^2 V \tag{10.48}$$

With these substitutions, (10.46) becomes

$$\oint\limits_S U\frac{\partial V}{\partial n}\,da=\int\limits_V\left[\nabla U\cdot\nabla V+U\nabla^2 V\right]dv \tag{10.49}$$

So far we haven't done much. Equation (10.49) is nothing more than the divergence theorem applied to the vector function $U\nabla V$. We can also write an equation similar to (10.49) where U and V are interchanged:

$$\oint\limits_S V\frac{\partial U}{\partial n}\,da=\int\limits_V\left[\nabla V\cdot\nabla U+V\nabla^2 U\right]dv \tag{10.50}$$

We subtract (10.50) from (10.49), which leads to (10.30) known as Green's theorem.

Exercises

Exercises for 10.1 Huygens' Principle as Formulated by Fresnel

P10.1 Huygens' principle can be used to describe refraction. Use a drawing program or a ruler and compass to produce a picture similar to Fig. 10.14, which shows that the graphical prediction of refracted angle from the Huygens' principle. Verify that the Huygens picture matches the numerical prediction from Snell's Law for an incident angle of your choice. Use $n_\mathrm{i} = 1$ and $n_\mathrm{t} = 2$.

HINT: Draw the wavefronts hitting the interface at an angle and treat each point where the wavefronts strike the interface as the source of circular waves propagating into the $n = 2$ material. The wavelength of the circular waves must be exactly half the wavelength of the incident light since $\lambda = \lambda_\mathrm{vac}/n$. Use at least four point sources and connect the matching wavefronts by drawing tangent lines as in the figure.

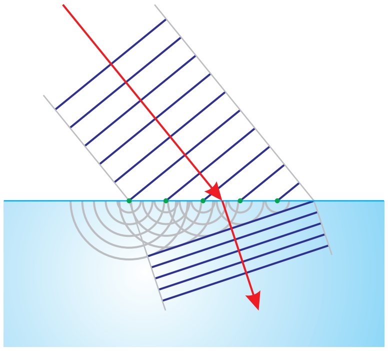


Figure 10.14

L10.2 (a) Why does the on-axis intensity behind a circular opening fluctuate (see Example 10.1) whereas the on-axis intensity behind a circular obstruction remains constant (see Example 10.2)?

(b) Create a collimated laser beam several centimeters wide. Observe the on-axis intensity on a movable screen (e.g. a hand-held card) behind a small circular aperture and behind a small circular obstruction placed in the beam. (video)

(c) In the case of the circular aperture, measure the distance to several on-axis minima and check that it agrees with (10.3).

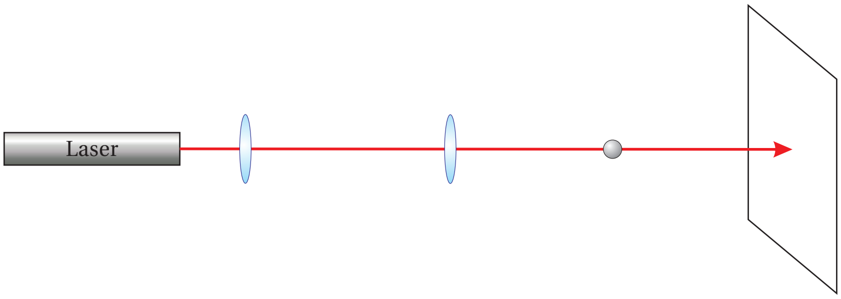


Figure 10.15

Exercises for 10.2 Scalar Diffraction Theory

P10.3 Show that $E(r) = E_0 r_0 e^{ikr}/r$ is a solution to the scalar Helmholtz equation (10.5).

HINT: In spherical coordinates

$$\nabla^2\psi = \frac{1}{r}\frac{\partial^2}{\partial r^2}\left(r\psi\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2}$$

P10.4 (a) A vector field is needed to satisfy Maxwell's equations instead of the scalar field in P10.3, whose real part after appending $e^{-i\omega t}$ is

$$E(r) = \frac{A}{r}\cos\left(kr - \omega t\right)$$

Let's attempt to create a vector field from this scalar field in the simplest way possible. From experience, we expect a transverse wave, which we take to oscillate in the $\hat{\phi}$ direction:

$$\mathbf{E}(r) = \frac{A}{r}\cos\left(kr - \omega t\right)\hat{\phi}$$

(i) Show that **E** satisfies Gauss's Law (1.1). (ii) Compute the curl of **E** in Faraday's Law (1.3) to deduce **B**. (iii) Show that this **B** satisfies Gauss' Law for magnetism (1.2). (iv) Finally, show that the above **E** and **B** do *not* satisfy Ampere's law (1.4).

HINT: In spherical coordinates

$$\nabla\cdot\mathbf{E} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2E_r\right) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta E_\theta\right) + \frac{1}{r\sin\theta}\frac{\partial E_\phi}{\partial\phi}$$

$$\nabla\times\mathbf{E} = \hat{\mathbf{r}}\frac{1}{r\sin\theta}\left[\frac{\partial}{\partial\theta}\left(\sin\theta E_\phi\right) - \frac{\partial E_\theta}{\partial\phi}\right] + \hat{\theta}\frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial E_r}{\partial\phi} - \frac{\partial}{\partial r}\left(rE_\phi\right)\right]$$
$$+\hat{\phi}\frac{1}{r}\left[\frac{\partial}{\partial r}\left(rE_\theta\right) - \frac{\partial E_r}{\partial\theta}\right]$$

(b) The following somewhat more complicated 'spherical' wave

$$\mathbf{E}(r,\theta) = \frac{A\sin\theta}{r}\left[\cos\left(kr - \omega t\right) - \frac{1}{kr}\sin\left(kr - \omega t\right)\right]\hat{\phi}$$

(i.e. the real part of (10.9) with time dependence appended) does satisfy Maxwell's equations. Describe how this wave behaves as a function of r and θ. What conditions need to be satisfied for this equation to be well approximated by the spherical wave in part (a)?

Exercises for 10.3 Fresnel Approximation

P10.5 By direct substitution, show that (10.16) satisfies the paraxial wave equation (10.15).

Exercises for 10.4 Fraunhofer Approximation

P10.6 Calculate the Fraunhofer diffraction *field* and *intensity* patterns for a rectangular aperture (dimensions Δx by Δy) illuminated by a plane wave E_0. In other words, derive (10.20).

P10.7 A single narrow slit has a mask placed over it so the aperture function is not a square profile but rather a cosine: $E(x', y', 0) = E_0 \cos(\pi x'/L)$ for $-L/2 < x' < L/2$ and $E(x', y', 0) = 0$ otherwise. Calculate the far-field (Fraunhofer) diffraction pattern. Make a plot of intensity as a function of $kLx/2z$; qualitatively compare the pattern to that of a regular single slit. Do not perform any integration in the y dimension. Write the intensity as being proportional to an x-dependent expression.

Exercises for 10.5 Diffraction with Cylindrical Symmetry

P10.8 (a) Repeat Example 10.1 to find the on-axis intensity (i.e. $\rho = 0$) after a circular aperture in both the Fresnel approximation (10.27) and the Fraunhofer approximation (10.28).

(b) Make suitable approximations directly to (10.3) to obtain the same answers as in part (a).

(c) Check how well the Fresnel and Fraunhofer approximations work by graphing the Fresnel- and Fraunhofer-approximation results together with (10.3) on a single plot as a function of z. Take $D = 10\ \mu$m and $\lambda = 500$ nm. To see the result better, use a log scale on the z-axis.

Answer:

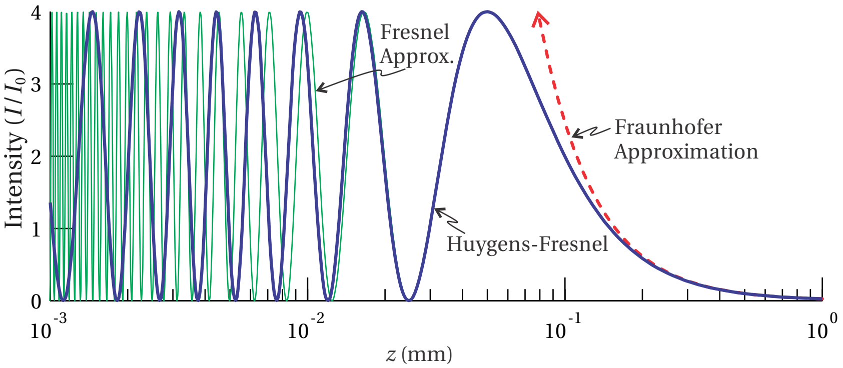


Figure 10.17 On-axis intensity behind a circular aperture calculated using the Fresnel diffraction formula (10.1), the Fresnel approximation (10.27), and the Fraunhofer approximation (10.28).

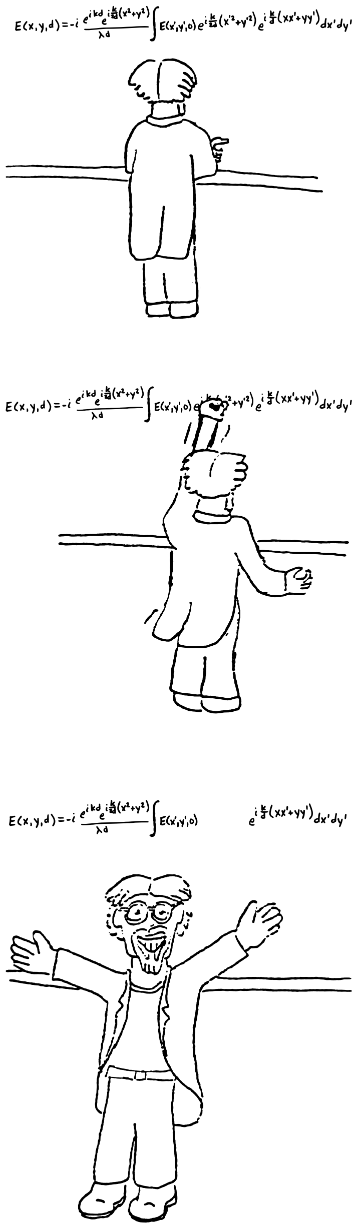

$E(x,y,d) = -i \frac{e^{ikd} e^{i\frac{k}{2d}(x^2+y^2)}}{\lambda d} \int E(x',y',0) e^{i\frac{k}{2d}(x'^2+y'^2)} e^{-i\frac{k}{d}(xx'+yy')} dx'dy'$
$E(x,y,d) = -i \frac{e^{ikd} e^{i\frac{k}{2d}(x^2+y^2)}}{\lambda d} \int E(x',y',0) e^{-i\frac{k}{d}(xx'+yy')} dx'dy'$

Figure 10.16 "The Fraunhofer Approximation" by Sterling Cornaby

P10.9 Calculate the Fraunhofer diffraction intensity pattern (10.29) for a circular aperture (diameter D) illuminated by a plane wave E_0. That is, repeat example 10.5 while filling in the integration step. For added benefit, try to do it without peeking.

Exercises for 10.A Fresnel-Kirchhoff Diffraction Formula

P10.10 Learn by heart the derivation of the Fresnel-Kirchhoff diffraction formula (outlined in Appendix 10.A). Indicate the percentage of how well you understand the derivation. If you write 100% percent, it means that you can reproduce the derivation without peeking.

Chapter 11

Diffraction Applications

In this chapter, we consider a number of practical examples of diffraction. We first discuss diffraction theory in systems involving lenses. The Fraunhofer diffraction pattern discussed in section 10.4, applicable in the far-field limit, is imaged to the focus of a lens when the lens is placed in the stream of light. This has important implications for the *resolution* of instruments such as telescopes or grating spectrometers.

The *array theorem*, which applies to the Fraunhofer limit, is introduced in section 11.3. This theorem is a powerful mathematical tool that enables one to deal conveniently with diffraction from an array of identical apertures. One of the important uses of the array theorem is in determining Fraunhofer diffraction from a grating, since a diffraction grating can be thought of as an array of narrow slit apertures. In section 11.5, we study the workings of a diffraction spectrometer and explore resolution limitations.

Finally, we consider a Gaussian laser beam to understand its focusing and diffraction properties. The information presented here comes up remarkably often in research activity. We often think of lasers as collimated beams of light that propagate indefinitely without expanding. However, the laws of diffraction require that every finite beam eventually grows in width. The rate at which a laser beam diffracts depends on its *beam waist* size. Because laser beams usually have narrow divergence angles and therefore obey the paraxial approximation, we can calculate their behavior via the Fresnel approximation discussed in section 10.3. Appendix 11.A discusses the ABCD law for Gaussian beams, which is a method of computing the effects of optical elements such as lenses on laser beams. The ABCD law arbitrates the competition between beam expansion via diffraction and beam focusing from traversing a lens.

11.1 Fraunhofer Diffraction with a Lens

The Fraunhofer limit corresponds to the ultimate amount of diffraction that light in an optical system experiences. As has been previously discussed, the Fraunhofer approximation applies to diffraction when the propagation distance from an aperture is sufficiently large (see (10.18) and (10.19)). Mathematically, it is obtained via a two-dimensional Fourier transform. The intensity of the far-field diffraction pattern is

$$I(x,y,z) = \frac{1}{2}c\epsilon_0 \left| \frac{1}{\lambda z} \iint\limits_{\text{aperture}} E(x',y',0)\, e^{-ik\left(\frac{x}{z}x' + \frac{y}{z}y'\right)} dx'\,dy' \right|^2 \tag{11.1}$$

Notice that the dependence of the diffraction on x, y, and z comes only through the combinations $\theta_x \cong x/z$ and $\theta_y \cong y/z$. Therefore, the diffraction pattern in the Fraunhofer limit is governed by the two angles θ_x and θ_y, and the pattern preserves itself indefinitely. As the light continues to propagate, the pattern increases in size at a rate proportional to distance traveled so that the *angular width* is preserved. The situation is depicted in Fig. 11.1.

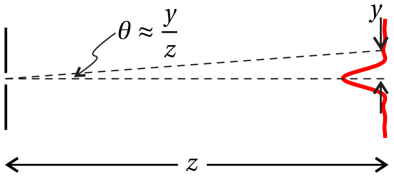


Figure 11.1 Diffraction in the far field.

Recall that in order to use the Fraunhofer diffraction formula we need to satisfy $z \gg \pi\left(\text{aperture radius}\right)^2/\lambda$ (see (10.18)). As an example, if an aperture with a 1 cm radius (not necessarily circular) is used with visible light, the light must travel more than a kilometer in order to reach the Fraunhofer limit. It may therefore seem unlikely to reach the Fraunhofer limit in a typical optical system, especially if the aperture or beam size is relatively large. Nevertheless, spectrometers, which typically utilize diffraction gratings many centimeters wide, depend on achieving the Fraunhofer limit within the confines of a manageable instrument box. This is accomplished using imaging techniques. The Fraunhofer limit is also naturally reached in other instruments that employ lenses such as telescopes.

Consider a lens with focal length f placed in the path of light following an aperture (see Fig. 11.2). Let the lens be placed an arbitrary distance L after the aperture. The lens produces an image of the Fraunhofer pattern at a new location d_i following the lens according to the imaging formula (see (9.56))

$$\frac{1}{f} = \frac{1}{-(z-L)} + \frac{1}{d_\text{i}}. \tag{11.2}$$

Keep in mind that the lens interrupts the light before the Fraunhofer pattern has a chance to form. This means that the Fraunhofer diffraction pattern may

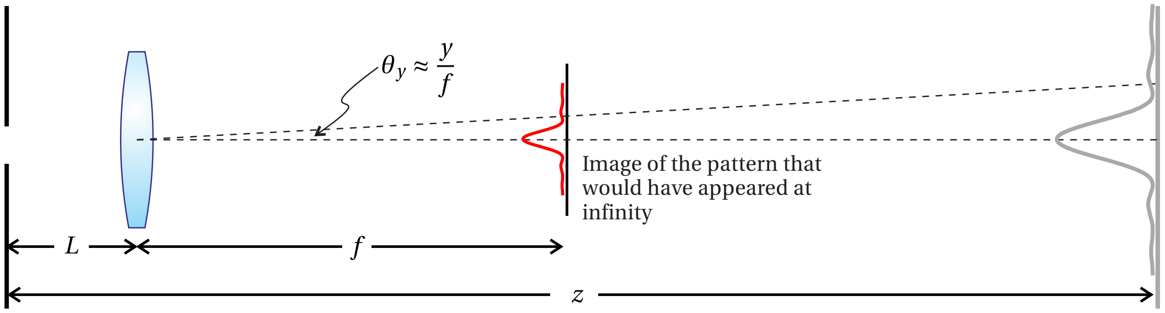

$\theta_y \approx \frac{y}{f}$
Image of the pattern that would have appeared at infinity
L
f
z

Figure 11.2 Imaging of the Fraunhofer diffraction pattern to the focus of a lens.

be thought of as a *virtual object* a distance $z-L$ to the right of the lens. Since the Fraunhofer diffraction pattern occurs at very large distances (i.e. $z\to\infty$) the image of the Fraunhofer pattern appears at the focus of the lens:

$$d_{\rm i}\cong f. \tag{11.3}$$

Thus, a lens makes it very convenient to observe the Fraunhofer diffraction pattern even from relatively large apertures. It is not necessary to let the light propagate for kilometers. We need only observe the pattern at the focus of the lens as shown in Fig. 11.2. Notice that the spacing L between the aperture and the lens is unimportant to this conclusion.

Even though we know that the Fraunhofer diffraction pattern occurs at the focus of a lens, the question remains as to the size of the image. To find the answer, let us examine the magnification (9.57), which is given by

$$M=-\frac{d_{\rm i}}{-(z-L)} \tag{11.4}$$

Taking the limit of very large z and employing (11.3), the magnification becomes

$$M\to\frac{f}{z} \tag{11.5}$$

This is a remarkable result. When the lens is inserted, the size of the diffraction pattern decreases by the ratio of the lens focal length f to the original distance z to a far-away screen. Since in the Fraunhofer regime the diffraction pattern is proportional to distance (i.e. size $\propto z$), the image at the focus of the lens scales in proportion to the focal length (i.e. size $\propto f$). This means that *the angular width of the pattern is preserved!* With the lens in place, we can rewrite (11.1) straightaway as

$$I\left(x,y,L+f\right)\cong\frac{1}{2}c\epsilon_0\left|\frac{1}{\lambda f}\iint\limits_{\text{aperture}}E\left(x',y',0\right)e^{-i\frac{k}{f}(xx'+yy')}\,dx'dy'\right|^2 \tag{11.6}$$

which describes the intensity distribution pattern at the focus of the lens.

Although (11.6) correctly describes the *intensity* at the focus of a lens, we cannot easily write the electric field since the imaging techniques that we have used do not easily render the phase information. To obtain an expression for the *field*, it will be necessary to employ the Fresnel diffraction formula, which we accomplish in the remainder of this section. Before doing so, we will need to know how a lens adjusts the phase fronts of the light passing through it.

Phase Front Alteration by a Lens

Consider a monochromatic light field that goes through a *thin* lens with focal length f. In traversing the lens, the wavefront undergoes a phase shift that varies across the lens. We will reference the phase shift to that experienced by the light

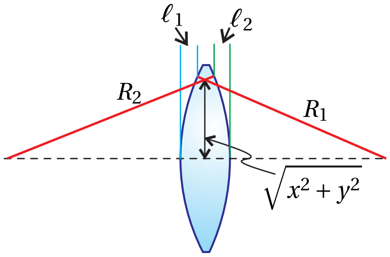


Figure 11.3 A thin lens, which modifies the phase of a field passing through.

that goes through the center of the lens. We take the distances ℓ_1 and ℓ_2, as drawn in Fig. 11.3, to be positive.

The light passing through the off-axis portion of the lens experiences less material than the light passing through the center. The difference in optical path length is $(n-1)(\ell_1+\ell_2)$ (see discussion connected with (9.13)). This means that the phase of the field passing through the off-axis portion of the lens relative to the phase of the field passing through the center is

$$\Delta\phi = -k(n-1)(\ell_1+\ell_2)\,. \tag{11.7}$$

The negative sign indicates a phase *advance* (i.e. same sign as $-\omega t$). In (11.7), k represents the wave number in vacuum (i.e. $2\pi/\lambda_{\text{vac}}$); since ℓ_1 and ℓ_2 correspond to distances outside of the lens material.

We can find expressions for ℓ_1 and ℓ_2 from the equations describing the spherical surfaces of the lens:

$$\begin{aligned}(R_1-\ell_1)^2+x^2+y^2 &= R_1^2\\(R_2+\ell_2)^2+x^2+y^2 &= R_2^2\end{aligned} \tag{11.8}$$

As drawn in Fig. 11.3, R_1 is a positive radius of curvature while R_2 is negative, in accordance with conventions in chapter 9. In the spirit of the Fresnel approximation, which takes place in the paraxial limit, it is appropriate to neglect the terms ℓ_1^2 and ℓ_2^2 in comparison to other terms present in (11.8). Within this approximation, the equations become

$$\ell_1 \cong \frac{x^2+y^2}{2R_1} \quad \text{and} \quad \ell_2 \cong -\frac{x^2+y^2}{2R_2} \tag{11.9}$$

Substitution into (11.7) yields

$$\Delta\phi = -k(n-1)\left(\frac{1}{R_1}-\frac{1}{R_2}\right)\frac{(x^2+y^2)}{2} = -\frac{k}{2f}\left(x^2+y^2\right) \tag{11.10}$$

where the focal length of a thin lens f has been introduced according to the lens-maker's formula (9.46).

In summary, the light traversing a lens experiences a relative phase shift given by

$$E\left(x,y,z_{\text{after lens}}\right) = E\left(x,y,z_{\text{before lens}}\right)e^{-i\frac{k}{2f}(x^2+y^2)} \tag{11.11}$$

Equation (11.11) introduces a wavefront curvature to the field. For example, if a plane wave (i.e. a uniform field E_0) passes through the lens, the field emerges with a spherical-like wavefront converging towards the focus of the lens.

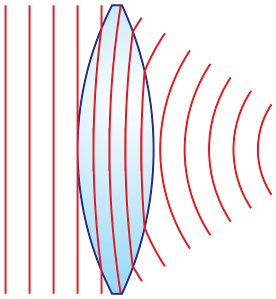


Figure 11.4 The phase fronts of a plane wave are bent as they pass through a lens.

We compute the diffraction pattern after the lens in three steps, as illustrated in Fig. 11.5. First, we use the Fresnel approximation to compute the field arriving at the lens. Second, we adjust the phase front of the light passing through the lens according to (11.11). Third, we use the field exiting the lens as the input for a second diffraction integral to find the field at the lens focus. The result gives an intensity pattern in agreement with (11.6) without ever employing the Fraunhofer approximation. It also provides the full expression for the field, including its phase.

Starting from the known field $E\left(x', y', 0\right)$ at the aperture, we compute the field incident on the lens using the Fresnel approximation:

$$E(x'', y'', L) = -i\frac{e^{ikL}e^{i\frac{k}{2L}\left(x''^2+y''^2\right)}}{\lambda L}\iint E(x', y', 0)e^{i\frac{k}{2L}\left(x'^2+y'^2\right)}e^{-i\frac{k}{L}\left(x''x'+y''y'\right)}dx'dy' \tag{11.12}$$

(The double primes keep track of distinct variables in sequential diffraction integrals.) As mentioned, the field gains a phase factor according to (11.11) upon transmitting through the lens. Finally, we use the Fresnel diffraction formula a second time to propagate the distance f from the back of the thin lens:

$$\begin{aligned} E\left(x, y, L+f\right) = -i\frac{e^{ikf}e^{i\frac{k}{2f}\left(x^2+y^2\right)}}{\lambda f}\iint &\left[E(x'', y'', L)e^{-i\frac{k}{2f}\left(x''^2+y''^2\right)}\right] \\ &\times e^{i\frac{k}{2f}\left(x''^2+y''^2\right)}e^{-i\frac{k}{f}\left(xx''+yy''\right)}dx''dy'' \end{aligned} \tag{11.13}$$

Right off, by choosing the propagation distance to be f, we get a nice cancellation of the phase factor introduced by the lens. Even so, as you can probably appreciate, the installation of (11.12) into (11.13) makes a rather long formula involving four dimensions of integration. Nevertheless, two of the integrals can be performed in advance of choosing the aperture (i.e. those over x'' and y''). This is accomplished with the help of the integral formula (0.55) (even though in this instance the real part of a is zero). After this cumbersome work, (11.13) reduces to

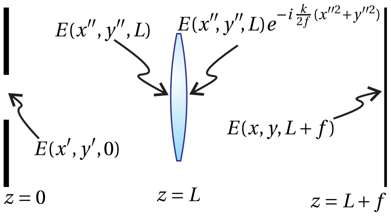


Figure 11.5 Diffraction from an aperture viewed at the focus of a lens.

$$E\left(x, y, L+f\right) = -i\frac{e^{ik(L+f)}e^{i\frac{k}{2f}\left(x^2+y^2\right)}e^{-i\frac{kL}{2f^2}\left(x^2+y^2\right)}}{\lambda f}\iint E(x', y', 0)e^{-i\frac{k}{f}\left(xx'+yy'\right)}dx'dy' \tag{11.14}$$

Notice that at least the integration portion of this formula looks exactly like the Fraunhofer diffraction formula! This happened even though in the preceding discussion we did not at any time specifically make the Fraunhofer approximation. The result (11.14) implies the intensity distribution (11.6) as anticipated. However, the phase of the field is also revealed in (11.14).

In general, the field carries a wavefront curvature as it passes through the focal plane of the lens. In the special case $L = f$, the diffraction formula takes a particularly simple form:

$$E(x', y', L+f)\big|_{L=f} = -i\frac{e^{2ikf}}{\lambda f}\iint E(x', y', 0)e^{-i\frac{k}{f}\left(xx'+yy'\right)}dx'dy' \tag{11.15}$$

When the lens is placed at this special distance following the aperture, the Fraunhofer diffraction pattern viewed at the focus of the lens carries a flat wavefront.

11.2 Resolution of a Telescope

In the previous section we learned that the Fraunhofer diffraction pattern appears at the focus of a lens. This has important implications for telescopes and other optical instruments. In essence, any optical instrument incorporates an aperture, limiting the light that enters. If nothing else, the diameter of a lens itself acts effectively as an aperture. The pupil of the human eye is an aperture that causes a Fraunhofer diffraction pattern to occur at the retina. Cameras have irises control the light entering the camera, again giving rise to a Fraunhofer diffraction pattern at the image plane.

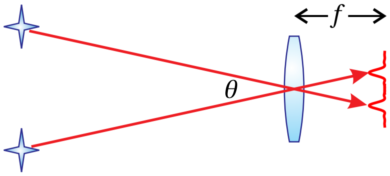


Figure 11.6 To resolve distinct images at the focus of a lens, the angular separation must exceed the width of the Fraunhofer diffraction patterns.

Of course, the focus of the lens is just where one wants to look in order to see images of distant objects. Of course, the Fraunhofer pattern, which occurs at the focus, represents the ultimate amount of diffraction caused by an aperture. This has the effect of blurring out features in the image and limiting *resolution*. This illustrates why it is impossible to focus light to a true point.

Suppose you point a telescope at two distant stars. An image of each star is formed in the focal plane of the lens. The angular separation between the two images (referenced from the lens) is the same as the angular separation between the stars.[1] This is depicted in Fig. 11.6. For reference, we are speaking of the image that forms between the objective lens and the eyepiece of a telescope, as seen in Fig. 9.15. Often, a CCD camera is placed at that image plane so that there is no need for an eyepiece.

A resolution problem occurs when the Fraunhofer diffraction pattern associated with each star causes them to blur by more than the angular separation between them. In this case the two images cannot be resolved because they 'bleed' into one another.

The Fraunhofer diffraction pattern from a circular aperture was computed previously (see (10.29)). At the focus of a lens, this pattern centered on each star becomes

$$I(\rho, f) = I_0 \left(\frac{\pi D^2}{4\lambda f}\right)^2 \left[2\frac{J_1\left(kD\rho/2f\right)}{\left(kD\rho/2f\right)}\right]^2 \tag{11.16}$$

where f, the focal length of the lens, takes the place of z in the diffraction formula. The parameter D is the diameter of the lens. This intensity pattern contains the first-order Bessel function J_1, which behaves somewhat like a sine wave as seen in Fig. 11.7. The main differences are that the zero crossings are not exactly periodic and the function slowly diminishes with larger arguments. The first zero crossing (after the origin) occurs at 1.22π.

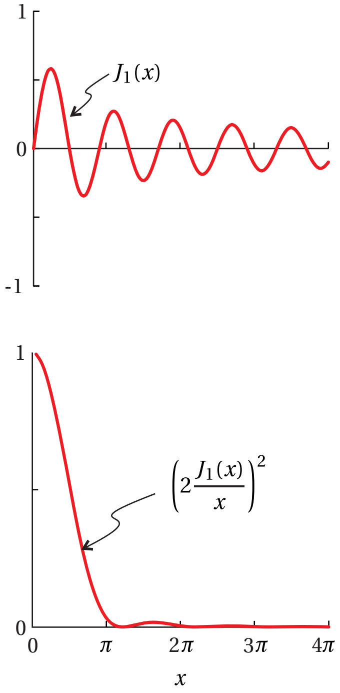


Figure 11.7 (a) First-order Bessel function. (b) Square of the Jinc function.

The intensity pattern described by (11.16) contains the factor $2J_1(\xi)/\xi$ (which

---

[1] In the thin-lens approximation, the ray from either star that traverses the center of the lens (i.e. $y = 0$) maintains its angle:

$$\begin{bmatrix} 0 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix} \begin{bmatrix} 0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} 0 \\ \theta_1 \end{bmatrix}$$

we call the jinc[2]), where ξ represents the combination $kD\rho/2f$. As noticed in Fig. 11.7, $J_1(\xi)$ goes to zero at $\xi = 0$. Thus, we have a zero-divided-by-zero situation similar to the sinc function (i.e. $\sin(\xi)/\xi$), which approaches one at the origin. The square of the jinc, shown in Fig. 11.7b, is proportional to the intensity described in (11.16). This pattern is sometimes called an Airy pattern after Sir George Biddell Airy (English, 1801–1892) who first described the pattern. As can be seen in Fig. 11.7b, the intensity quickly drops at larger radii.

We now return to the question of whether the images of two nearby stars as depicted in Fig. 11.6 can be distinguished. Since the peak in Fig. 11.7b is the dominant feature in the diffraction pattern, we will say that the two stars are resolved if the angle between them is enough to keep the diffracted versions of their images from seriously overlapping. We will adopt the criterion suggested by Lord Rayleigh that the peaks are distinguishable if the peak of one pattern is no closer than the first zero to the other peak. This situation is shown in Fig. 11.8.

The angle that corresponds to this separation of diffraction patterns is found by setting the argument of (11.16) equal to 1.22π, the location of the first zero:

$$\frac{kD\rho}{2f} = 1.22\pi \tag{11.17}$$

With a little rearranging we have

$$\theta_{\text{min}} \cong \frac{\rho}{f} = \frac{1.22\lambda}{D} \tag{11.18}$$

Here we have associated the ratio ρ/f (i.e. the radius of the diffraction pattern compared to the distance from the lens) with an angle θ_{min}. Again, the angle between the images (referenced from the lens) and the angle between the objects is the same. The *Rayleigh criterion* requires that the diffraction patterns be separated by at least this angle before we say that they are resolved.

θ_{min} depends on the diameter of the lens D as well as on the wavelength of the light. This analysis assumes that the light from the two objects is *incoherent*, meaning the intensities in the image plane add; interferences between the two fields fluctuate rapidly in time and average away.

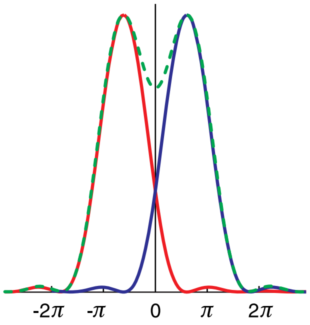


Figure 11.8 The Rayleigh criterion for a circular aperture.

Example 11.1

What minimum telescope-lens diameter is required to distinguish a Jupiter-like planet (orbital radius 8×10^8 km) from its star if they are 10 light-years away?

Solution: From (11.18) and assuming 500 nm light, we need

$$D > \frac{1.22\lambda}{\theta_{\text{min}}} = \frac{1.22(500 \times 10^{-9}\text{m})}{(8 \times 10^{11}\text{m})/(10\text{ly})} \times \frac{9.5 \times 10^{15}\text{m}}{\text{ly}} = 0.07\text{m}$$

This seems like a piece of cake; a telescope with a diameter bigger than 7cm will do the trick. However, the vastly unequal brightness of the star and the planet is

---

[2] often defined without the factor of 2

> the real technical challenge. The diffraction rings in the star's diffraction pattern completely swamp the faint signal from the planet.

11.3 The Array Theorem

In this section we develop the array theorem, which is used for calculating the Fraunhofer diffraction from an array of N identical apertures. We will be using the theorem to compute diffraction from a grating, which may be thought of as a mask with many closely spaced identical slits. However, the array theorem can be applied to apertures with any shape and configuration, as suggested by Fig. 11.9.

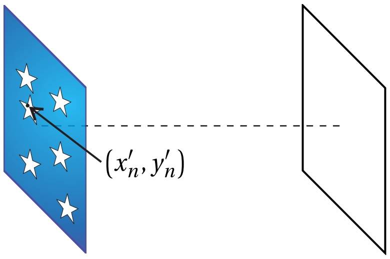


Figure 11.9 Array of identical apertures.

Consider N apertures in a mask, each with the identical field distribution described by $E_{\text{aperture}}(x', y', 0)$. Each identical aperture has a unique location on the mask. Let the location of the n^{th} aperture be designated by the coordinates (x'_n, y'_n). The field associated with the n^{th} aperture is then $E_{\text{aperture}}(x'-x'_n, y'-y'_n, 0)$, where the offset in the arguments shifts the location of the aperture. The field comprising all of the identical apertures is

$$E\left(x', y', 0\right) = \sum_{n=1}^{N} E_{\text{aperture}}(x' - x'_n, y' - y'_n, 0) \tag{11.19}$$

We next compute the Fraunhofer diffraction pattern for the above field. Upon inserting (11.19) into the Fraunhofer diffraction formula (10.19) we obtain

$$E\left(x, y, z\right) = -i\frac{e^{ikz} e^{i\frac{k}{2z}\left(x^2+y^2\right)}}{\lambda z} \sum_{n=1}^{N} \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' E_{\text{aperture}}\left(x' - x'_n, y' - y'_n, 0\right) e^{-i\frac{k}{z}\left(xx'+yy'\right)} \tag{11.20}$$

where we have taken the summation out in front of the integral. We have also integrated over the entire (infinitely wide) mask, taking E_{aperture} to be zero except inside each aperture.

Even without yet choosing the shape of the identical apertures, we can make some progress on (11.20) with the change of variables $x'' \equiv x' - x'_n$ and $y'' \equiv y' - y'_n$:

$$E\left(x, y, z\right) = -i\frac{e^{ikz} e^{i\frac{k}{2z}\left(x^2+y^2\right)}}{\lambda z} \sum_{n=1}^{N} \int_{-\infty}^{\infty} dx'' \int_{-\infty}^{\infty} dy'' E_{\text{aperture}}\left(x'', y'', 0\right) \times e^{-i\frac{k}{z}\left[x\left(x''+x'_n\right)+y\left(y''+y'_n\right)\right]} \tag{11.21}$$

Next we pull the factor $\exp\{-i\frac{k}{z}(xx'_n + yy'_n)\}$ out in front of the integral to arrive at our final result:

$$E\left(x,y,z\right)=\left[\sum_{n=1}^{N}e^{-i\frac{k}{z}\left(xx'_n+yy'_n\right)}\right]$$
$$\times\left[-i\frac{e^{ikz}e^{i\frac{k}{2z}\left(x^2+y^2\right)}}{\lambda z}\int_{-\infty}^{\infty}dx'\int_{-\infty}^{\infty}dy'E_{\text{aperture}}\left(x',y',0\right)e^{-i\frac{k}{z}\left(xx'+yy'\right)}\right] \tag{11.22}$$

For the sake of elegance, we have traded back x' for x'' and y' for y'' as the variables of integration. Equation (11.22) is known as the array theorem.[3] Note that the second factor in brackets is exactly the Fraunhofer diffraction pattern from a single aperture centered on $x'=0$ and $y'=0$. When more than one identical aperture is present, we only need to evaluate the Fraunhofer diffraction formula for a single aperture. Then, the single-aperture result is multiplied by the summation in front, which contains entirely the information about the placement of the multiple identical apertures.

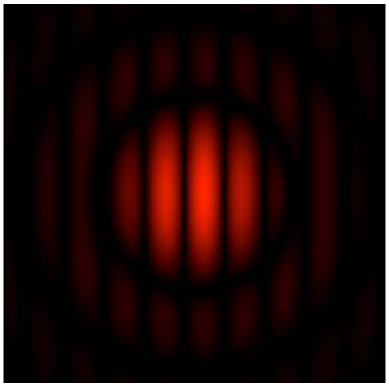


Figure 11.10 Fraunhofer diffraction pattern from two identical circular holes separated by twice their diameters.

Example 11.2

Calculate the Fraunhofer diffraction pattern for two identical circular apertures with diameter D whose centers are separated by a spacing h.

Solution: As computed previously, the single-slit Fraunhofer diffraction pattern from a circular aperture is given by (10.29). This is multiplied by (the square of) the factor on the first line of the array theorem (11.22), which gives an overall intensity pattern of

$$I\left(x,y,z\right)=\left|\sum_{n=1}^{2}e^{-i\frac{k}{z}\left(xx'_n+yy'_n\right)}\right|^2\times I_0\left(\frac{\pi D^2}{4\lambda z}\right)^2\left[2\frac{J_1\left(kD\rho/2z\right)}{\left(kD\rho/2z\right)}\right]^2$$

Let $y'_1=y'_2=0$. To create the separation h, let $x'_1=-h/2$ and $x'_2=h/2$. Then

$$\sum_{n=1}^{2}e^{-i\frac{k}{z}\left(xx'_n+yy'_n\right)}=e^{-i\frac{k}{z}\left(-\frac{hx}{2}\right)}+e^{-i\frac{k}{z}\left(\frac{hx}{2}\right)}=2\cos\left(\frac{khx}{2z}\right)$$

---

[3] A somewhat abstract alternative route to the array theorem recognizes that the field for each aperture can be written as a 2-D convolution (see P0.26) between the aperture function $E_{\text{aperture}}\left(x',y',0\right)$ and delta functions specifying the aperture location:

$$E_{\text{aperture}}\left(x'-x'_n,y'-x'_n,0\right)=\int_{-\infty}^{\infty}dx'\int_{-\infty}^{\infty}dy'\delta\left(x''-x'_n\right)\delta\left(y''-x'_n\right)E_{\text{aperture}}\left(x'-x'',y'-y'',0\right)$$

The integral in (11.20) therefore may be viewed as a 2-D Fourier transform of a convolution, where kx/z and ky/z play the role of *spatial frequencies*. The convolution theorem (see P0.26) indicates that this is the same as the product of Fourier transforms. The 2-D Fourier transform for the delta function (times 2π) is

$$\int_{-\infty}^{\infty}dx''\int_{-\infty}^{\infty}dy''\delta\left(x''-x'_n\right)\delta\left(y''-y'_n\right)e^{-i\frac{k}{z}\left(xx''+yy''\right)}=e^{-i\frac{k}{z}\left(xx'_n+yy'_n\right)}$$

The array theorem (11.22) exhibits this factor. It multiplies the single-slit Fraunhofer diffraction integral, which is the Fourier transform of the other function.

The overall pattern then becomes

$$I(x,y,z) = I_0\left(\frac{\pi D^2}{2\lambda z}\right)^2\left[2\frac{J_1(kD\rho/2z)}{(kD\rho/2z)}\right]^2\cos^2\left(\frac{khx}{2z}\right)$$

This pattern can be seen in Fig. 11.10.

11.4 Diffraction Grating

In this section we will use the array theorem to calculate the Fraunhofer diffraction from a grating comprised of an array of equally spaced identical slits. An array of uniformly spaced slits is called a transmission grating (see Fig. 11.11). Reflection gratings are similar, being composed of an array of narrow rectangular mirrors that behave similarly to the slits.

Let the slit apertures be positioned at

$$x'_n = \left(n - \frac{N+1}{2}\right)h, \qquad y'_n = 0 \tag{11.23}$$

where N is the total number of slits. Then the summation in the array theorem, (11.22), becomes

$$\sum_{n=1}^{N} e^{-i\frac{k}{z}\left(xx'_n+yy'_n\right)} = e^{i\frac{khx}{z}\left(\frac{N+1}{2}\right)}\sum_{n=1}^{N} e^{-i\frac{khx}{z}n} \tag{11.24}$$

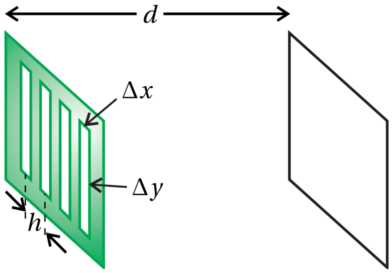


Figure 11.11 Transmission grating.

This summation is recognized as a geometric sum, which can be performed using formula (0.65). Equation (11.24) then simplifies to

$$\begin{aligned}\sum_{n=1}^{N} e^{-i\frac{k}{z}\left(xx'_n+yy'_n\right)} &= e^{i\frac{k}{z}\left(\frac{N+1}{2}\right)xh}e^{-i\frac{khx}{z}}\frac{e^{-i\frac{khx}{z}N}-1}{e^{-i\frac{khx}{z}}-1}\\ &= \frac{e^{-i\frac{khx}{2z}N}-e^{i\frac{khx}{2z}N}}{e^{-i\frac{khx}{2z}}-e^{i\frac{khx}{2z}}} = \frac{\sin\left(N\frac{khx}{2z}\right)}{\sin\left(\frac{khx}{2z}\right)}\end{aligned} \tag{11.25}$$

The diffraction pattern for a single slit was previously calculated in example 10.4. When (11.25) and (10.20) are installed in the array theorem (11.22), we get for the intensity

$$I(x,y,z) = \frac{\sin^2\left(N\frac{khx}{2z}\right)}{\sin^2\left(\frac{khx}{2z}\right)}\left[I_0\frac{\Delta x^2\Delta y^2}{\lambda^2 z^2}\operatorname{sinc}^2\left(\frac{\pi\Delta x}{\lambda z}x\right)\operatorname{sinc}^2\left(\frac{\pi\Delta y}{\lambda z}y\right)\right] \tag{11.26}$$

This is the Fraunhofer diffraction pattern for the overall grating.

The y dependence in (11.26) is typically unimportant in applications where spectral information is revealed in the x-dimension only. Moreover, the incident field often does not have a uniform strength along the entire slit in the

y-dimension, making the diffraction pattern along the y dimension different from $\text{sinc}\left[\left(\pi\Delta y/\lambda z\right)y\right]$ anyway. Since y is of little relevance, we can consider the pattern in (11.26) for fixed y, say $y = 0$. The intensity pattern in the horizontal dimension may be written as

$$I(x) = I_{\text{peak}}\text{sinc}^2\left(\frac{\pi\Delta x}{\lambda z}x\right)\frac{\sin^2\left(N\frac{\pi hx}{\lambda z}\right)}{N^2\sin^2\left(\frac{\pi hx}{\lambda z}\right)} \tag{11.27}$$

Note that $\lim\limits_{\alpha\to 0}\frac{\sin N\alpha}{\sin\alpha} = N$ so we have placed N^2 in the denominator and absorbed the same factor into the definition of I_{peak}, which represents the intensity on the screen at $x = 0$. Again, the intensity I_{peak} is associated with a given value of y.

It is left as an exercise to study the functional form of (11.27), especially how the number of slits N influences the behavior. The case of $N = 2$ describes the diffraction pattern for a Young's double slit experiment. We now have a description of the Young's two-slit pattern in the case that the slits have finite openings of width Δx rather than infinitely narrow ones.

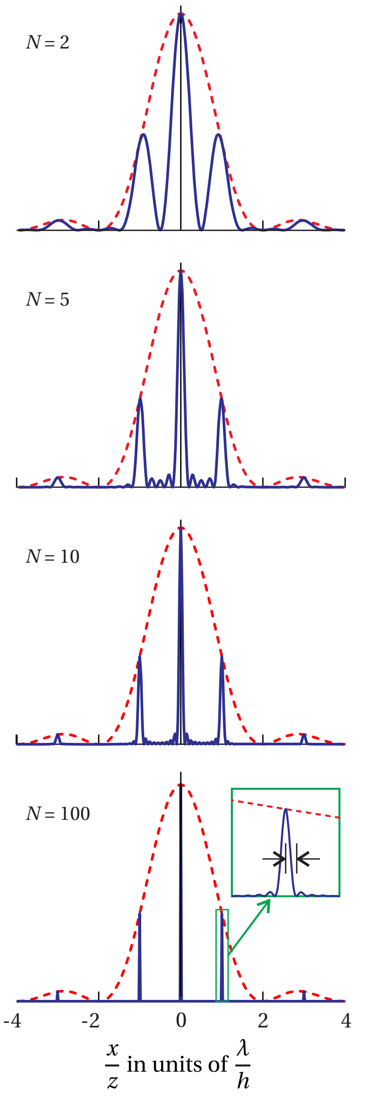


Figure 11.12 Plot of the intensity at a screen for diffraction through various numbers of slits, each with $\Delta x = h/2$ (slit widths half the separation). The vertical scale is arbitrary and different for each plot (assuming the same illumination). The dotted line shows the single slit diffraction pattern. (a) Diffraction from a double slit. (b) Diffraction from 5 slits. (c) Diffraction from 10 slits. (d) Diffraction from 100 slits with an inset illustrating the width of a peak.

11.5 Spectrometers

The formula (11.27) can be exploited to make wavelength measurements. This forms the basis of a diffraction grating spectrometer. In order to achieve good spatial separation between wavelengths, it is necessary to allow the light to propagate a far distance. Optimal wavelength separation therefore occurs in the Fraunhofer regime for which (11.27) applies.

A spectrometer has relatively poor resolving power compared to a Fabry-Perot interferometer. Nevertheless, a spectrometer is not hampered by the serious limitation imposed by free spectral range. A spectrometer is able to measure a wide range of wavelengths simultaneously. The Fabry-Perot interferometer and the grating spectrometer in this sense are complementary, the one being able to make very precise measurements within a narrow wavelength range and the other being able to characterize wide ranges of wavelengths simultaneously.

To appreciate how a spectrometer works, consider Fraunhofer diffraction from a grating, as described by (11.27). The structure of the diffraction pattern has various peaks. For example, Fig. 11.12a shows the diffraction peaks from a Young's double slit (i.e. $N = 2$). The diffraction pattern is comprised of the typical Young's double-slit pattern multiplied by the diffraction pattern of a single slit. (Note that $\sin^2\left(2\frac{\pi hx}{\lambda z}\right)/4\sin^2\left(\frac{\pi hx}{\lambda z}\right) = \cos^2\left(\frac{\pi hx}{\lambda z}\right)$.)

As the number of slits N increases, the peaks tend to sharpen while staying in the same location as the peaks in the Young's double-slit pattern. Figure 11.12b shows the case for $N = 5$. The prominent peaks occur when $\sin(\pi hx/\lambda z)$ in the denominator of (11.27) goes to zero. Keep in mind that the numerator goes to zero at the same places, creating a zero-over-zero situation, so the peaks are not infinitely tall.

With larger values of N, the peaks can become extremely sharp, and the small secondary peaks in between become tiny in comparison. Fig. 11.12c shows the case of $N = 10$ and Fig. 11.12d, shows the case of $N = 100$.

When very many slits are used, the resulting sharp diffraction peaks becomes very useful for measuring spectra of light, since the position of the diffraction peaks depends on wavelength (except for the center peak at $x = 0$). If light of different wavelengths is simultaneously present, then the diffraction peaks associated with different wavelengths appear in different locations.

Consider the inset in Fig. 11.12d, which gives a close-up view of the first-order diffraction peak for $N = 100$. The location of this peak on a distant screen varies with the wavelength of the light. How much must the wavelength change to cause the peak to move by half of its 'width' as marked in the inset of Fig. 11.12d? This corresponds to the minimum wavelength separation that allows two associated peaks to be distinguished.

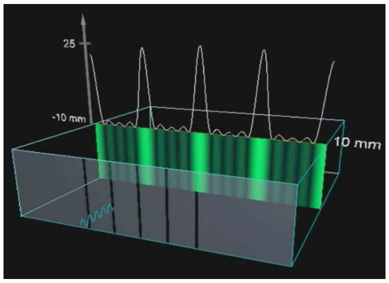


Figure 11.13 Animation showing diffraction through a number of slits.

Finding the Minimum Distinguishable Wavelength Separation

As mentioned, the main diffraction peaks occur when the denominator of (11.27) goes to zero, i.e.

$$\frac{\pi h x}{\lambda z} = m\pi \tag{11.28}$$

The numerator of (11.27) goes to zero at these same locations (i.e. $N\pi h x/\lambda z = Nm\pi$), so the peaks remain finite. If two nearby wavelengths λ_1 and λ_2 are sent through the grating simultaneously, their m^{th} peaks are located at

$$x_1 = \frac{mz\lambda_1}{h} \quad \text{and} \quad x_2 = \frac{mz\lambda_2}{h} \tag{11.29}$$

These are spatially separated by

$$\Delta x_\lambda \equiv x_2 - x_1 = \frac{mz}{h}\Delta\lambda \tag{11.30}$$

where $\Delta\lambda \equiv \lambda_2 - \lambda_1$.

Meanwhile, we can find the spatial width of, say, the first peak by considering the change in x_1 that causes the sine in the numerator of (11.27) to reach the nearby zero (see inset in Fig. 11.12d). This condition implies

$$N\frac{\pi h\left(x_1 + \Delta x_{\text{peak}}\right)}{\lambda_1 z} = Nm\pi + \pi \tag{11.31}$$

We will say that two peaks, associated with λ_1 and λ_2, are barely distinguishable when $\Delta x_\lambda = \Delta x_{\text{peak}}$. We also substitute from (11.29) to rewrite (11.31) as

$$N\frac{\pi h\left(mz\lambda_1/h + mz\Delta\lambda/h\right)}{\lambda_1 z} = Nm\pi + \pi \quad \Rightarrow \quad \Delta\lambda = \frac{\lambda}{Nm} \tag{11.32}$$

Here we have dropped the subscript on the wavelength in the spirit of $\lambda_1 \approx \lambda_2 \approx \lambda$.

As we did for the Fabry-Perot interferometer, we can define the *resolving power* of the diffraction grating as

$$RP \equiv \frac{\lambda}{\Delta\lambda} = mN \tag{11.33}$$

The resolving power is proportional to the number of slits illuminated on the diffraction grating. The resolving power also improves for higher diffraction orders m.

Example 11.3

What is the resolving power with $m = 1$ of a 2-cm-wide grating with 500 slits per millimeter, and how wide is the 1st-order diffraction peak for 500-nm light after 1-m focusing?

Solution: From (11.33) the resolving power is

$$RP = mN = 2\text{ cm}\frac{500}{0.1\text{ cm}} = 10^4$$

and the minimum distinguishable wavelength separation is

$$\Delta\lambda = \lambda/RP = 500\text{ nm}/10^4 = 0.05\text{ nm}$$

From (11.30), with $z \to f$, we have

$$\Delta x = \frac{mf}{h}\Delta\lambda = \frac{1\text{ m}}{2\times 10^{-6}\text{ m}}0.05\text{nm} = 25\ \mu\text{m}$$

As illustrated in the previous example, it is common to employ a focusing optic to reach the Fraunhofer limit within a convenient distance. In addition, since the array theorem requires the same illumination of each slit, the incident light should be collimated or plane-wave like. This is also accomplished using a lens. Figure 11.14 illustrates the typical layout. Light enters a narrow slit located at the focus of a concave mirror. The first mirror collimates the light, and the collimated light then strikes a reflective diffraction grating. The first-order diffracted light then travels to a second concave mirror which focuses the diffracted light to an exit slit, where the Fraunhofer diffraction pattern of the grating appears. If a CCD camera is positioned at the focus to record many wavelengths at once, the instrument is called a spectrometer. If instead an exit slit is placed at the focus so that only one wavelength at a time emerges through the slit, the instrument is called a monochromator. In the latter case, the angle of the grating can be scanned to cause different wavelengths to transmit through the exit slit.

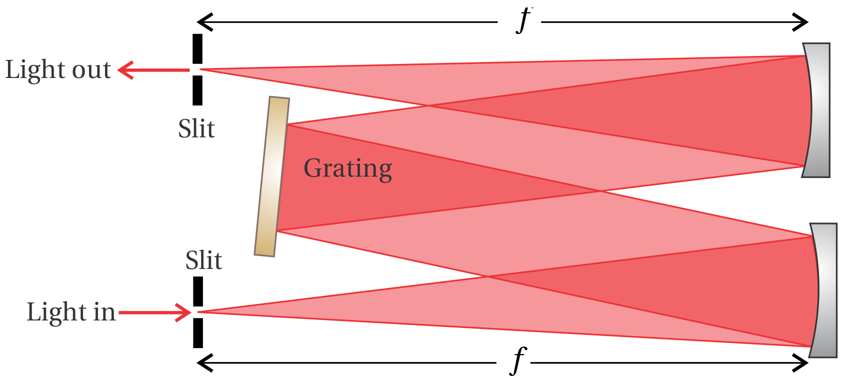


Figure 11.14 Symmetric monochromator layout.

11.6 Diffraction of a Gaussian Field Profile

Consider a Gaussian field profile (in the plane $z = 0$) described with the functional form

$$E(x', y', 0) = E_0 e^{-\frac{x'^2+y'^2}{w_0^2}} \tag{11.34}$$

The parameter w_0 is called the *beam waist*, which specifies the radius of Gaussian profile. It is depicted in Fig. 11.15. To better appreciate the meaning of w_0, consider the intensity of the above field distribution:

$$I\left(x', y', 0\right) = I_0 e^{-2\rho'^2/w_0^2} \tag{11.35}$$

where $\rho'^2 \equiv x'^2 + y'^2$. In (11.35) we see that w_0 indicates the radius at which the intensity reduces by the factor $e^{-2} = 0.135$.

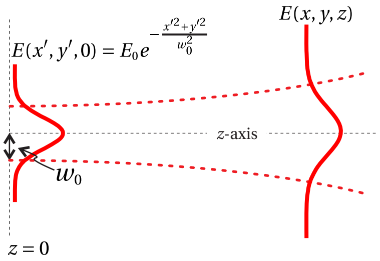


Figure 11.15 Diffraction of a Gaussian field profile.

We would like to know how this field evolves when it propagates forward from the plane $z = 0$. Notice that the phase of (11.34) is uniform or plane-wave like. We therefore expect the beam to expand outward as it diffracts along z.[4] We compute the field downstream using the Fresnel approximation (10.13):

$$E\left(x, y, z\right) = -i\frac{e^{ikz}e^{i\frac{k}{2z}\left(x^2+y^2\right)}}{\lambda z}\int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' \left[E_0 e^{-\left(x'^2+y'^2\right)/w_0^2}\right] e^{i\frac{k}{2z}\left(x'^2+y'^2\right)} e^{-i\frac{k}{z}\left(xx'+yy'\right)} \tag{11.36}$$

The Gaussian profile itself limits the dimension of the 'aperture', so there is no problem with integrating to infinity. Equation (11.36) can be rewritten as

$$E\left(x, y, z\right) = -i\frac{E_0 e^{ikz}e^{i\frac{k}{2z}\left(x^2+y^2\right)}}{\lambda z}\int_{-\infty}^{\infty} dx' e^{-\left(\frac{1}{w_0^2}-i\frac{k}{2z}\right)x'^2-i\frac{kx}{z}x'} \int_{-\infty}^{\infty} dy' e^{-\left(\frac{1}{w_0^2}+i\frac{k}{2z}\right)y'^2-i\frac{ky}{z}y'} \tag{11.37}$$

---

[4] The beam would converge to narrower widths if instead we used a phase associated with converging wavefronts like those on the left of Fig. 11.17.

The integrals over x' and y' have the identical form and can be done individually with the help of the integral formula (0.55). The algebra is cumbersome, but the integral in the x' dimension becomes

$$\begin{aligned}\int_{-\infty}^{\infty} dx' e^{-\left(\frac{1}{w_0^2}-i\frac{k}{2z}\right)x'^2-i\frac{kx}{z}x'} &= \left(\frac{\pi}{\frac{1}{w_0^2}-i\frac{k}{2z}}\right)^{\frac{1}{2}} \exp\left(\frac{\left(-i\frac{kx}{z}\right)^2}{4\left(\frac{1}{w_0^2}-i\frac{k}{2z}\right)}\right)\\ &= \left(\frac{\pi}{-i\frac{k}{2z}\left(1+i\frac{2z}{kw_0^2}\right)}\right)^{\frac{1}{2}} \exp\left(\frac{-kx^2}{2z\left(\frac{2z}{kw_0^2}-i\right)}\right)\\ &= \left(\frac{\lambda z}{\sqrt{1+\left(\frac{2z}{kw_0^2}\right)^2}\, e^{i\tan^{-1}\frac{2z}{kw_0^2}}}\right)^{\frac{1}{2}} \exp\left(\frac{-kx^2\left[\frac{2z}{kw_0^2}+i\right]}{2z\left[1+\left(\frac{2z}{kw_0^2}\right)^2\right]}\right)\end{aligned} \tag{11.38}$$

A similar expression results from the integration on y'.

When (11.38) and the equivalent expression for the y-dimension are used in (11.37), the result is

$$E(x,y,z) = E_0 \frac{e^{ikz}e^{i\frac{k}{2z}(x^2+y^2)}}{\sqrt{1+\left(\frac{2z}{kw_0^2}\right)^2}}\, e^{-\frac{(x^2+y^2)}{1+\left(\frac{2z}{kw_0^2}\right)^2}\left(\frac{1}{w_0^2}+i\frac{k}{2z}\right)}\, e^{-i\tan^{-1}\frac{2z}{kw_0^2}} \tag{11.39}$$

This rather complicated-looking expression for the field distribution is in fact very useful and can be directly interpreted, as discussed in the next section.

Gaussian Field in Cylindrical Coordinates

A Gaussian field profile is one of few diffraction problems that can be handled conveniently in either the Cartesian (as above) or cylindrical coordinate. In cylindrical coordinates, the Fresnel diffraction integral (10.27) is

$$E(\rho,z) = -\frac{2\pi i e^{ikz}e^{i\frac{k\rho^2}{2z}}}{\lambda z}\int_0^{\infty}\rho' d\rho' E_0 e^{-\rho'^2/w_0^2}e^{i\frac{k\rho'^2}{2z}} J_0\left(\frac{k\rho\rho'}{z}\right)$$

We can use the integral formula (0.59) to obtain

$$E\left(\rho,z\right) = -iE_0 \frac{2\pi e^{ikz} e^{i\frac{k\rho^2}{2z}}}{\lambda z} \frac{e^{-\frac{\left(\frac{k\rho}{z}\right)^2}{4\left[\frac{1}{w_0^2} - i\frac{k}{2z}\right]}}}{2\left[\frac{1}{w_0^2} - i\frac{k}{2z}\right]}$$

$$= E_0 \frac{e^{ikz} e^{i\frac{k\rho^2}{2z}}}{\sqrt{1+\left(\frac{2z}{kw_0^2}\right)^2}} e^{-\frac{\rho^2}{1+\left(\frac{2z}{kw_0^2}\right)^2}\left(\frac{1}{w_0^2}+i\frac{k}{2z}\right)} e^{-i\tan^{-1}\frac{2z}{kw_0^2}}$$

which is identical to (11.39).

11.7 Gaussian Laser Beams

The cumbersome Gaussian-field expression (11.39) can be cleaned up through judicious introduction of new quantities:

$$E\left(\rho,z\right) = E_0 \frac{w_0}{w\left(z\right)} e^{-\frac{\rho^2}{w^2(z)}} e^{ikz + i\frac{k\rho^2}{2R(z)} - i\tan^{-1}\frac{z}{z_0}} \tag{11.40}$$

where

$$\rho^2 \equiv x^2 + y^2, \tag{11.41}$$

$$w\left(z\right) \equiv w_0\sqrt{1 + z^2/z_0^2}, \tag{11.42}$$

$$R\left(z\right) \equiv z + z_0^2/z, \tag{11.43}$$

$$z_0 \equiv \frac{kw_0^2}{2} \tag{11.44}$$

This formula describes the lowest-order Gaussian mode, the most common *laser beam* profile.[5]

It turns out that (11.40) works equally well for negative values of z. The expression can therefore be used to describe the field of a simple laser beam everywhere (before and after it goes through a focus). In fact, the expression works also near $z = 0$! At $z = 0$ the diffracted field (11.40) returns the exact expression for the original field profile (11.34) (see P11.11).There is good reason for this since the Fresnel diffraction integral is an exact solution to the paraxial wave equation (10.15). The beam (11.40) satisfies the paraxial wave equation for positive and negative z. In short, (11.40) may be used with impunity as long as the divergence angle of the beam is not too wide.

---

[5] Lasers can also be *multimode*, exhibiting more complicated structure through 'higher-order' modes.

As we analyze (11.40), consider the intensity profile $I \propto |E|^2$ as depicted in Fig. 11.16:

$$I\left(\rho,z\right) = I_0 \frac{w_0^2}{w^2(z)} e^{-\frac{2\rho^2}{w^2(z)}} = \frac{I_0}{1+z^2/z_0^2} e^{-\frac{2\rho^2}{w^2(z)}} \tag{11.45}$$

By inspection, we see that $w(z)$ gives the radius of the beam anywhere along z. At $z = 0$, the *beam waist*, $w(z=0)$ reduces to w_0, as expected. The parameter z_0, known as the *Rayleigh range*, specifies the distance along the axis from $z = 0$ to the point where the intensity decreases by a factor of 2. Note that w_0 and z_0 are not independent of each other but are connected through the wavelength according to (11.44). There is a tradeoff: a small beam waist means a short *depth of focus*. That is, a small w_0 means a small Rayleigh range z_0.

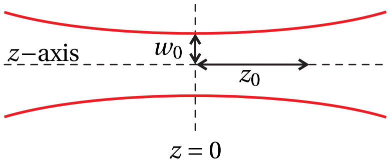


Figure 11.16 A Gaussian laser field profile in the vicinity of its beam waist.

We next consider the phase terms that appear in the field expression (11.40). The phase term $ikz+ik\rho^2/2R(z)$ describes the phase of curved wavefronts, where $R(z)$ is the radius of curvature of the wavefront at z. The curvature of wavefronts is evident in Fig. 11.17. At $z = 0$, the radius of curvature is infinite (see (11.43)), meaning that the wavefront is flat at the laser beam waist. In contrast, at very large values of z we have $R(z) \cong z$ (see (11.43)). In this case, we may write these phase terms as $kz + \frac{k\rho^2}{2z} \cong k\sqrt{z^2+\rho^2}$. This describes a spherical wavefront emanating from the origin out to point $\left(\rho,z\right)$. The Fresnel approximation represents spherical wavefronts as parabolic curves (same as the paraxial approximation). As a reminder, to restore the temporal dependence of the field, we append $e^{-i\omega t}$ to the solution, as discussed in connection with (10.4).

The phase $-i\tan^{-1} z/z_0$ is perhaps a bit more mysterious. It is called the *Gouy shift* and is actually present for any light that goes through a focus, not just laser beams. The Gouy shift is not overly dramatic since the expression $\tan^{-1} z/z_0$ ranges from $-\pi/2$ (at $z=-\infty$) to $\pi/2$ (at $z=+\infty$). Nevertheless, when light goes through a focus, it experiences an overall phase shift of π.

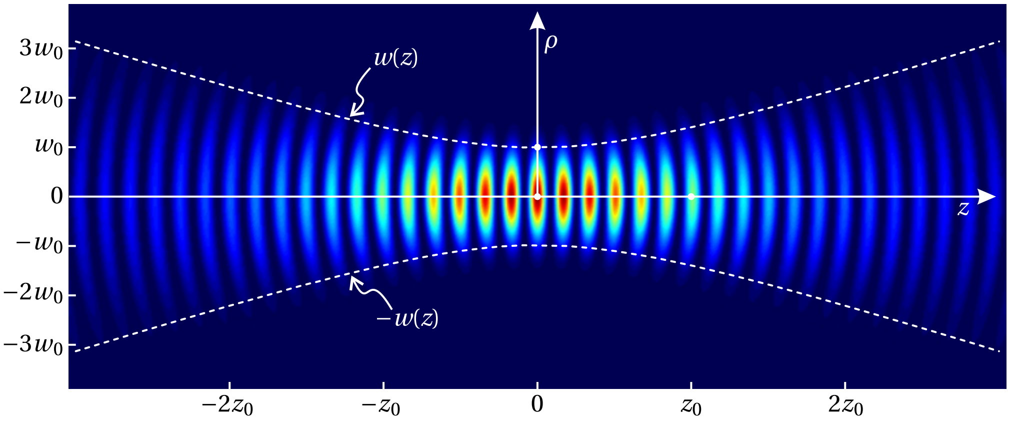


Figure 11.17 Real part of a Gaussian laser field at an instant in time. The radius of curvature of wavefronts is apparent.

Example 11.4

Write the beam waist w_0 in terms of the *f-number*, defined to be the ratio of z to the beam diameter $2w(z)$ far from the beam waist.

Solution: Far away from the beam waist (i.e. $z >> z_0$) the laser beam expands along a cone. That is, its diameter increases in proportion to distance.

$$w(z) = w_0\sqrt{1+z^2/z_0^2} \to w_0 z/z_0$$

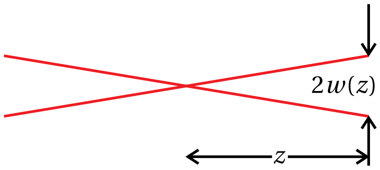


Figure 11.18

The cone angle is parameterized by the f-number, the ratio of the cone height to its base:

$$f^\# \equiv \lim_{z\to\pm\infty} \frac{z}{2w(z)} = \frac{z}{2w_0 z/z_0} = \frac{z_0}{2w_0}$$

Substitution of (11.44) into this expression yields

$$w_0 = \frac{2\lambda f^\#}{\pi} \tag{11.46}$$

Equation (11.46) gives a convenient way to predict the size of a laser focus. One calculates the f-number by dividing the diameter of the beam far from the focus into the distance from the focus. In practice you may be very surprised at how poorly a beam may focus in comparison with the theoretical prediction (due to aberrations). It is always good practice to directly measure your focus if its size is important to an experiment.

Appendix 11.A ABCD Law for Gaussian Beams

In this section we discuss and justify the ABCD law for Gaussian beams. The law enables one to predict the parameters of a Gaussian beam that exits from an optical system, given the parameters of an input Gaussian beam. To make the prediction, one needs only the ABCD matrix for the optical system, taken as a whole. The system may be arbitrarily complex with many optical components.

At first, it may seem unlikely that such a prediction should be possible since ABCD matrices were introduced to describe the propagation of rays. On the other hand, Gaussian beams are governed by the laws of diffraction. As an example of this dichotomy, consider a collimated Gaussian beam that traverses a converging lens. By ray theory, one expects the Gaussian beam to focus near the focal point of the lens. However, a collimated beam by definition is already in the act of going through focus. In the absence of the lens, there is a tendency for the beam to grow via diffraction, especially if the beam waist is small. This tendency competes with the focusing effect of the lens, and a new beam waist can occur at a wide range of locations, depending on the exact outcome of this competition.

A Gaussian beam is characterized by its Rayleigh range z_0. From this, the beam waist radius w_0 may be extracted via (11.44), assuming the wavelength is known. Suppose that a Gaussian beam encounters an optical system at position

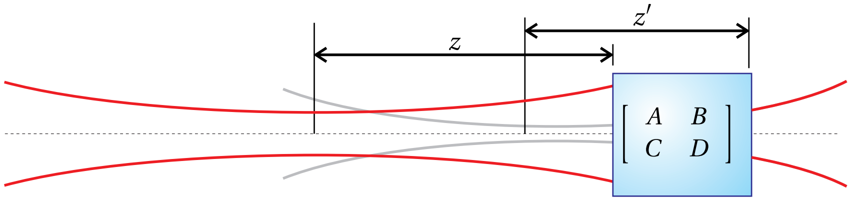


Figure 11.19 Gaussian laser beam traversing an optical system described by an ABCD matrix. The dark lines represent the incoming and exiting beams. The gray line represents where the exiting beam *appears* to have been.

z, referenced to the position of the beam's waist as shown in Fig. 11.19. The beam exiting from the system, in general, has a new Rayleigh range z_0'. The waist of the new beam also occurs at a different location. Let z' denote the location of the exit of the optical system, referenced to the location of the waist of the new beam. If the exiting beam diverges as in Fig. 11.19, then it emerges from a *virtual* beam waist located before the exit point of the system. In this case, z' is taken to be positive. On the other hand, if the emerging beam converges to an actual waist, then z' is taken to be negative since the exit point of the system occurs before the focus.

The ABCD law is embodied in the following relationship:[6]

$$z' + iz_0' = \frac{A(z+iz_0)+B}{C(z+iz_0)+D} \tag{11.47}$$

where A, B, C, and D are the matrix elements of the optical system. The imaginary number $i \equiv \sqrt{-1}$ imbues the law with complex arithmetic. It makes two equations from one, since the real and imaginary parts of (11.47) must separately be equal.

We now prove the ABCD law. We begin by showing that the law holds for two specific ABCD matrices. First, consider the matrix for propagation through a distance d:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \tag{11.48}$$

We know that simple propagation has minimal effect on a beam. The Rayleigh range is unchanged, so we expect that the ABCD law should give $z_0' = z_0$. The propagation through a distance d modifies the beam position by $z' = z + d$. We now check that the ABCD law agrees with these results by inserting (11.48) into (11.47):

$$z' + iz_0' = \frac{1(z+iz_0)+d}{0(z+iz_0)+1} = z+d+iz_0 \quad \text{(propagation through distance d)} \tag{11.49}$$

Thus, the law holds in this case.

Next we consider the ABCD matrix of a thin lens (or a curved mirror):

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix} \tag{11.50}$$

---

[6] The complex conjugate of this expression works equally well.

A beam that traverses a thin lens undergoes the phase shift $-k\rho^2/2f$, according to (11.11). This modifies the original phase of the wavefront $k\rho^2/2R(z)$, seen in (11.40). The phase of the exiting beam is therefore

$$\frac{k\rho^2}{2R(z')} = \frac{k\rho^2}{2R(z)} - \frac{k\rho^2}{2f} \tag{11.51}$$

where we do not keep track of unimportant overall phases such as kz or kz'. With (11.43) this relationship reduces to

$$\frac{1}{R(z')} = \frac{1}{R(z)} - \frac{1}{f} \Rightarrow \frac{1}{z' + z'^2_0/z'} = \frac{1}{z + z_0^2/z} - \frac{1}{f} \tag{11.52}$$

In addition to this relationship, the local radius of the beam given by (11.42) cannot change while traversing the 'thin' lens. Therefore,

$$w\left(z'\right) = w(z) \Rightarrow z_0'\left(1 + \frac{z'^2}{z'^2_0}\right) = z_0\left(1 + \frac{z^2}{z_0^2}\right) \tag{11.53}$$

On the other hand, the ABCD law for the thin lens gives

$$z' + iz_0' = \frac{1(z + iz_0) + 0}{-\left(1/f\right)(z + iz_0) + 1} \qquad \text{(traversing a thin lens with focal length f)} \tag{11.54}$$

It is left as an exercise (see P11.14) to show that (11.54) is consistent with (11.52) and (11.53).

So far we have shown that the ABCD law works for two specific examples, namely propagation through a distance d and transmission through a thin lens with focal length f. From these elements we can derive more complicated systems. However, the ABCD matrix for a thick lens cannot be constructed from just these two elements. We can construct the matrix for a thick lens if we sandwich a thick *window* (as opposed to empty space) between two thin lenses (see P9.9). The proof that the matrix for a thick window obeys the ABCD law is left as an exercise (see P11.17). With these relatively few elements, essentially any optical system can be constructed, provided that the beam propagation begins and ends in the same index of refraction.

To complete our proof of the general ABCD law, we need only show that when it is applied to the compound element

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix} \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} = \begin{bmatrix} A_2A_1 + B_2C_1 & A_2B_1 + B_2D_1 \\ C_2A_1 + D_2C_1 & C_2B_1 + D_2D_1 \end{bmatrix} \tag{11.55}$$

it gives the same answer as when the law is applied sequentially, first on

$$\begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix}$$

and then on

$$\begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix}$$

Explicitly, we have

$$
\begin{aligned}
z'' + iz_0'' &= \frac{A_2\left(z' + iz_0'\right) + B_2}{C_2\left(z' + iz_0'\right) + D_2} \\
&= \frac{A_2\left[\frac{A_1(z+iz_0)+B_1}{C_1(z+iz_0)+D_1}\right] + B_2}{C_2\left[\frac{A_1(z+iz_0)+B_1}{C_1(z+iz_0)+D_1}\right] + D_2} \\
&= \frac{A_2\left[A_1(z+iz_0)+B_1\right] + B_2\left[C_1(z+iz_0)+D_1\right]}{C_2\left[A_1(z+iz_0)+B_1\right] + D_2\left[C_1(z+iz_0)+D_1\right]} \\
&= \frac{(A_2A_1 + B_2C_1)(z+iz_0) + (A_2B_1 + B_2D_1)}{(C_2A_1 + D_2C_1)(z+iz_0) + (C_2B_1 + D_2D_1)} \\
&= \frac{A(z+iz_0)+B}{C(z+iz_0)+D}
\end{aligned}
\tag{11.56}
$$

Thus, we can construct any ABCD matrix that we wish from matrices that are known to obey the ABCD law. The resulting matrix also obeys the ABCD law.

Exercises

Exercises for 11.1 Fraunhofer Diffraction with a Lens

P11.1 Fill in the steps leading to (11.14) starting from (11.12) and (11.13). Show that the intensity distribution (11.6) is consistent with (11.14).

L11.2 Set up a collimated 'plane wave' in the laboratory using a HeNe laser ($\lambda = 633$ nm) and appropriate lenses.

(a) Choose a rectangular aperture (Δx by Δy) and place it in the plane wave. Observe the Fraunhofer diffraction on a very far away screen (i.e. where $z \gg \frac{k}{2}\left(\text{aperture radius}\right)^2$ is satisfied). Check that the location of the 'zeros' agrees with (10.20).

(b) Place a lens in the beam after the aperture. Use a CCD camera to observe the Fraunhofer diffraction profile at the focus of the lens. Check that the location of the 'zeros' agrees with (10.20), replacing z with f.

(c) Repeat parts (a) and (b) using a circular aperture with diameter D. Check the position of the first 'zero'. (video)

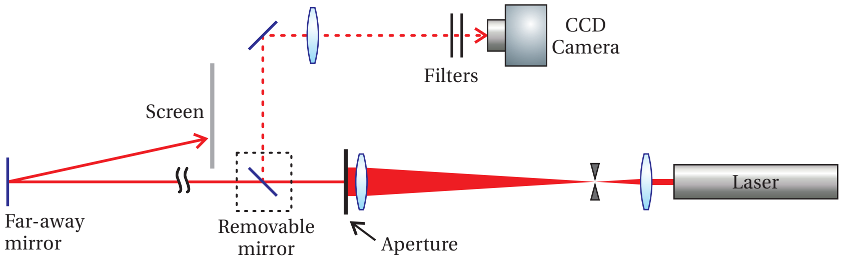


Figure 11.20

Exercises for 11.2 Resolution of a Telescope

P11.3 On the night of April 18, 1775, a signal was sent from the Old North Church steeple to Paul Revere, who was 1.8 miles away: "One if by land, two if by sea." If in the dark, Paul's pupils had 4 mm diameters, what is the minimum possible separation between the two lanterns that would allow him to correctly interpret the signal? Assume that the predominant wavelength of the lanterns was 580 nm.

HINT: You don't need to worry about refractive index inside the eye, $n = 1.33$. This causes the angular separation between the images to be $\theta/1.33$ inside the eye. The wavelength also shortens to 580 nm/1.33, causing a smaller diffraction pattern. As far as resolution is concerned, the two effects exactly compensate.

L11.4 Simulate two stars using laser beams ($\lambda = 633$ nm). Align them nearly parallel with a small lateral displacement. (A mirror can aid in getting the beams very close.) Send the beams down a long corridor until diffraction causes both beams to blend together so that it is no longer apparent that they are from two distinct sources. Use a lens to image the two sources onto a CCD camera. Use a variable iris near the lens to create different diameters.

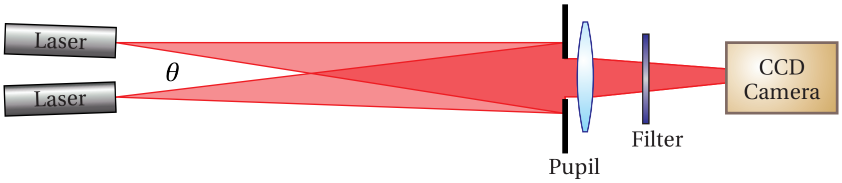


Figure 11.21

Experimentally determine the diameter D that just allows you to resolve the two sources according to the Rayleigh criterion. Check your measurement against theoretical prediction. (video)

HINT: The angular separation between the two sources is obtained by dividing the lateral separation of the beams into the propagation distance.

Exercises for 11.3 The Array Theorem

P11.5 Find the Fraunhofer diffraction pattern created by an array of nine circles, each with radius a, which are centered at the following (x', y') coordinates: $(-b, b)$, $(0, b)$, (b, b), $(-b, 0)$, $(0, 0)$, $(b, 0)$, $(-b, -b)$, $(0, -b)$, $(b, -b)$. Make a plot of the result for the situation where $b = 5a$. Scale position with units of $z/(ka)$. View the plot at different 'zoom levels' to see both course and fine detail.

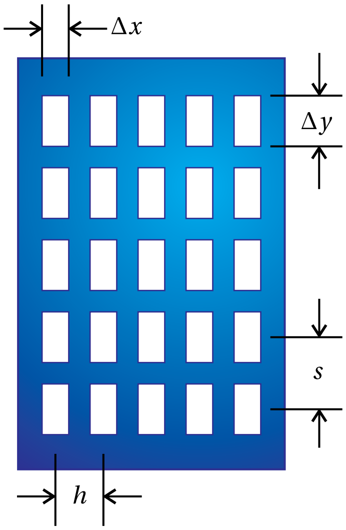


Figure 11.22

P11.6 (a) A plane wave is incident on a screen of N^2 uniformly spaced identical rectangular apertures of dimension Δx by Δy (see Fig. 11.22). Their positions are described by $x_n = h\left(n - \frac{N+1}{2}\right)$ and $y_m = s\left(m - \frac{N+1}{2}\right)$. Find the far-field (Fraunhofer) pattern of the light transmitted by the grid.

(b) You look at a distant sodium street lamp (somewhat monochromatic) through a curtain made from a fine mesh fabric with crossed threads. Make a sketch of what you expect to see (how the lamp will look to you).

HINT: Remember that the lens of your eye causes the Fraunhofer diffraction of the mesh to appear at the retina.

Exercises for 11.4 Diffraction Grating

P11.7 Consider Fraunhofer diffraction from a grating of N slits having widths Δx and equal separations h. Make plots (label relevant points and scaling) of the intensity pattern for $N = 1$, $N = 2$, $N = 5$, and $N = 1000$ in the case where $h = 2\Delta x$, $\Delta x = 5\ \mu$m, and $\lambda = 500$ nm. Let the Fraunhofer diffraction be observed at the focus of a lens with focal length $f = 100$ cm. Do you expect I_{peak} to be the same value for all of these cases?

Exercises for 11.5 Spectrometers

P11.8 For the case of $N = 1000$ in P11.7, you wish to position a narrow slit at the focus of the lens so that it transmits only the first-order diffraction peak (i.e. at $khx/(2f) = \pm\pi$). (a) How wide should the slit be if it is to match the width of the peak (as defined in (11.31))?

(b) What small change in wavelength (away from $\lambda = 500$ nm) will cause the intensity peak to shift by the width of the slit found in part (a)?

L11.9 (a) Use a HeNe laser to determine the period h of a reflective grating.

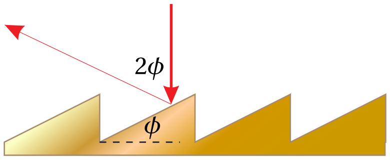


Figure 11.23

(b) Give an estimate of the *blaze angle* ϕ on the grating. HINT: Assume that the blaze angle is optimized for first-order diffraction of the HeNe laser (for one side) at normal incidence. The blaze angle enables a mirror-like reflection of the diffracted light on each groove. (video)

(c) You have two mirrors of focal length 75 cm and the reflective grating in the lab. You also have two very narrow adjustable slits and the ability to 'tune' the angle of the grating. Sketch how to use these items to make a monochromator. If the beam that hits the grating is 5 cm wide, what do you expect the ultimate resolving power of the monochromator to be in the wavelength range of 500 nm? HINT: See Fig. 11.14.

L11.10 Study the Jarrell Ash monochromator. Use a tungsten lamp as a source and observe how the instrument works by taking the entire top off. Do not breathe-on or touch the optical surfaces when you do this. In the dark, trace the light inside of the instrument with a card and observe what happens when you change the wavelength setting. Place the top back on when you are done. (video)

(a) Predict the best theoretical resolving power that this instrument can do assuming 1200 lines per millimeter. You will need to measure the width of the grating, which is matched well to the width of each mirror.

(b) What should the width Δx of the entrance and exit slits be to obtain this resolving power? Assume $\lambda = 500$ nm.

HINT: Set Δx to the peak width of a given wavelength, as defined in (11.31).

(c) Using the width of the mirrors, determine the f-number (defined in example 11.4) of the monochromator. Ideally, an external lens will couple light into the monochromatic using the same f-number. Compute the beam waist (11.46) for a Gaussian beam and check whether it approximately matches the size of the slit found in part (b).

Exercises for 11.7 Gaussian Laser Beams

P11.11 (a) Confirm that (11.40) reduces to (11.34) when $z = 0$.

(b) Take the limit $z \gg z_0$ of (11.40) to find the far-field form of the beam, which is the Fraunhofer diffraction of the laser focus.

P11.12 Use the Fraunhofer integral formula (either (10.19) or (10.28)) to determine the far-field pattern of a Gaussian laser focus (11.34).

HINT: The answer should agree with P11.11 part (b).

L11.13 Consider the following setup where a diverging laser beam is collimated using an uncoated lens. A double reflection from the two surfaces of the lens (known as a ghost) comes out in the forward direction, focusing after a short distance. Use a CCD camera to study this focused beam. The collimated beam serves as a reference to reveal the phase of the focused beam through interference. Because the weak ghost beam concentrates near its focus, the two beams can have similar intensities for optimal fringe visibility. (video)[7]

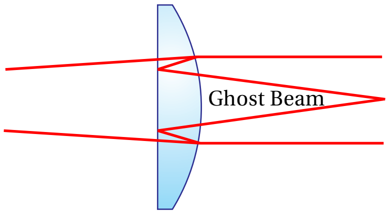


Figure 11.24

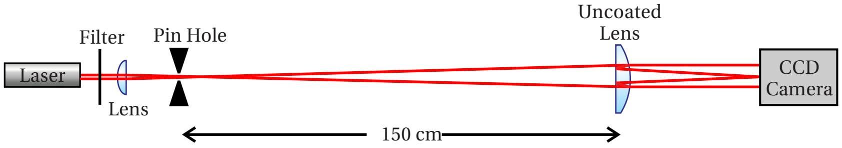


Figure 11.25

The ghost beam $E_1(\rho, z)$ is described by (11.40), where the origin is at the focus. Let the collimated beam be approximated as a plane wave $E_2 e^{ikz+i\phi}$, where ϕ is the relative phase between the two beams. The net intensity is then $I_t(\rho, z) \propto \left|E_1(\rho, z) + E_2 e^{ikz+i\phi}\right|^2$ or

$$I_t(\rho, z) = \left[I_2 + I_1(\rho, z) + 2\sqrt{I_2 I_1(\rho, z)} \cos\left(\frac{k\rho^2}{2R(z)} - \tan^{-1}\frac{z}{z_0} - \phi\right)\right]$$

---

[7] J. Peatross and M. V. Pack, “Viewing the Mathematical Structure of Gaussian Laser Beams in a Student Laboratory,” Am. J. Phys. **69**, 1169 (2001).

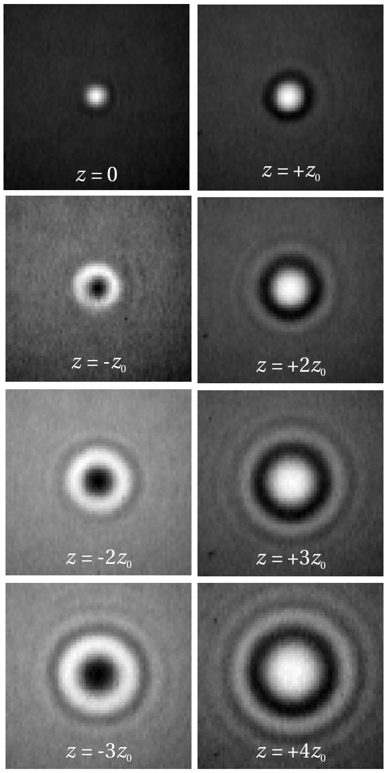


Figure 11.26

where $I_1(\rho, z)$ is given by (11.45). We now have a formula that retains both $R(z)$ and the Gouy shift $\tan^{-1} z/z_0$, which are not present in the intensity distribution of a single beam (see (11.45)).

(a) Determine the f-number for the ghost beam (see example 11.4). Use this measurement to predict a value for w_0. HINT: You know that at the lens, the focusing beam is the same size as the collimated beam.

(b) Measure the actual spot size w_0 at the focus. How does it compare to the prediction?

HINT: Before measuring the spot size, make a subtle adjustment to the tilt of the lens. This incidentally causes the phase between the two beams to vary by small amounts, which you can set to $\phi = \pm\pi/2$. Then *at the focus* the cosine term vanishes and the two beams don't interfere (i.e. the intensities simply add). This is accomplished if the center of the interference pattern is as dark as possible either far before or far after the focus.

(c) Observe the effect of the Gouy shift. Since $\tan^{-1} z/z_0$ varies over a range of π, you should see that the ring pattern inverts as you move the camera from before the focus to after the focus (i.e. the bright rings exchange with the dark ones).

(d) Predict the Rayleigh range z_0 and check that the radius of curvature $R(z) \equiv z + z_0^2/z$ agrees with measurement at a small distance from the focus.

HINT: You should see interference rings similar to those in Fig. 11.26. The only phase term that varies with ρ is $k\rho^2/2R(z)$. If you count N fringes out to a radius ρ, then $k\rho^2/2R(z)$ has varied by $2\pi N$.

Exercises for 11.A ABCD Law for Gaussian Beams

P11.14 Find the solutions to (11.54) (i.e. find z' and z_0' in terms of z and z_0). Show that the results are in agreement with (11.52) and (11.53).

P11.15 Assuming a collimated beam (i.e. $z = 0$ and beam waist w_0), find the location $L = -z'$ and size w_0' of the subsequent focus when the beam goes through a thin lens with focal length f.

L11.16 Place a long-focal-length lens (e.g. $f = 100$ cm) in a HeNe laser beam soon after the exit mirror of the cavity where the beam waist w_0 is sub millimeter. Characterize the focus of the resulting laser beam using filters and a CCD camera, and compare the results with the expressions derived in P11.15.

P11.17 Prove the ABCD law for a beam propagating through a thick window of material with matrix

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & d/n \\ 0 & 1 \end{bmatrix}$$

Chapter 12

Interferograms and Holography

In chapter 8, we studied a Michelson interferometer in an idealized sense: 1) The light entering the instrument was considered to be a plane wave. 2) The retro-reflecting mirrors were considered to be aligned perpendicular to the beams impinging on them. 3) All reflective surfaces were taken to be perfectly flat. If any of these conditions are not met, the beam emerging from the interferometer is likely to exhibit an interference or *fringe pattern*. A recorded fringe pattern (for example, on a CCD camera) is called an *interferogram*. In this chapter, we examine typical fringe patterns that can be produced in an interferometer. Such patterns are very useful for testing the prescription and quality of optical components.[1]

We will also study *holography*, where an interference pattern (or fringe pattern) is recorded and then later used to diffract light, in much the same way that gratings diffract light.[2] A recorded fringe pattern, when used for this purpose, is called a *hologram*. When light diffracts from a hologram, it can mimic the light field originally used to generate the fringe pattern. This is true even for complicated fields, recorded when light scatters from arbitrary three-dimensional objects. When the light field is re-created through diffraction, the resulting image looks 'three-dimensional', since the holographic fringes re-construct the original light field over a wide range of viewing angles.

12.1 Interferograms

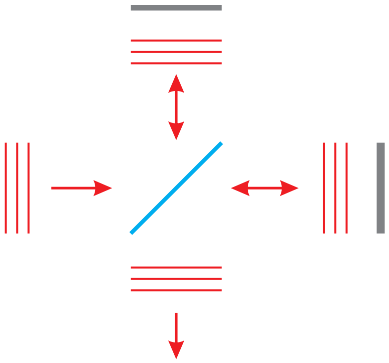


Figure 12.1 Michelson interferometer.

Consider the Michelson interferometer seen in Fig. 12.1. Suppose that the beam-splitter divides the fields evenly, so that the overall output intensity is given by (8.1):

$$I_{\text{tot}} = 2I_0\left[1+\cos(\omega\tau)\right] \tag{12.1}$$

As a reminder, τ is the roundtrip delay time of one path relative to the other. This equation is based on the idealized case, where the amplitude and phase of the two beams are uniform and perfectly aligned to each other following the beamsplitter. The entire beam 'blinks' on and off as the delay path τ is varied.

---

[1] See M. Born and E. Wolf, *Principles of Optics*, 7th ed., Sect. 7.5.5 (Cambridge: Cambridge University Press, 1999).

[2] In fact, a grating can be considered to be a hologram and holographic techniques are often employed to produce gratings.

What happens if one of the retro-reflecting mirrors is misaligned by a small angle θ? The fringe patterns seen in Fig. 12.2 (a)-(c) are the result. By the law of reflection, the beam returning from the misaligned mirror deviates from the 'ideal' path by an angle 2θ. This puts a relative phase variation of

$$\phi = kx\sin(2\theta_x) + ky\sin\left(2\theta_y\right) \tag{12.2}$$

on the misaligned beam.[3] Here θ_x represents the tilt of the mirror in the x-dimension and θ_y represents the amount of tilt in the y-dimension.

When the two plane waves join, the resulting intensity pattern is

$$I_{\text{tot}} = 2I_0\left[1 + \cos\left(\phi + \omega\tau\right)\right] \tag{12.3}$$

The phase term ϕ depends on the local position within the beam through x and y. Regions of uniform phase, called *fringes* (in this case individual stripes), have the same intensity. As the delay τ is varied, the fringes seem to 'move' across the detector. In this case, the fringes appear at one edge of the beam and disappear at the other.

Another interesting situation arises when the beams in a Michelson interferometer are diverging. A fringe pattern of concentric circles will be seen at the detector when the two beam paths are unequal (see Fig. 12.2 (d)). The radius of curvature for the beam traveling the longer path is increased by the added amount of delay $d = \tau c$. Thus, if beam 1 has radius of curvature R_1 when returning to the beam splitter, then beam 2 will have radius $R_2 = R_1 + d$ upon return (assuming flat mirrors). The relative phase (see phase term in (11.40)) between the two beams is

$$\phi = k\rho^2/2R_1 - k\rho^2/2R_2 \tag{12.4}$$

and the intensity pattern at the detector is given as before by (12.3).

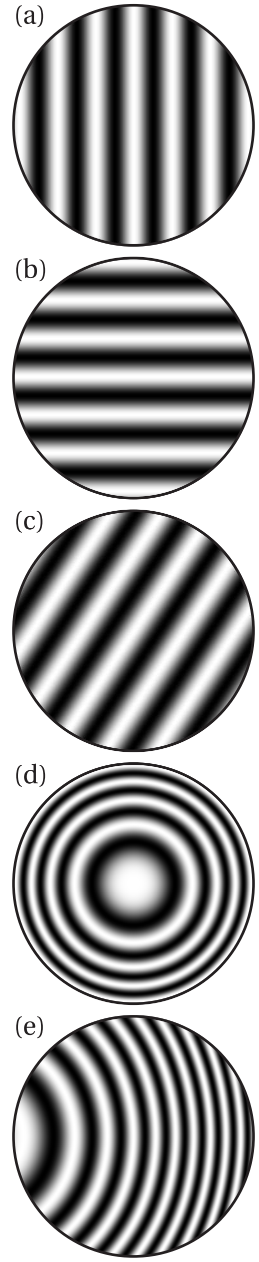


Figure 12.2 Fringe patterns for a Michelson interferometer: (a) Horizontally misaligned beams. (b) Vertically misaligned beams. (c) Both vertically and horizontally misaligned beams. (d) Diverging beam with unequal paths. (e) Diverging beam with unequal paths and horizontal misalignment.

12.2 Testing Optical Surfaces

A Michelson interferometer is ideal for testing the quality of optical surfaces. If any of the flat surfaces (including the beam splitter) in the interferometer are distorted, the fringe pattern readily reveals it. Figure 12.3 shows an example of a fringe pattern when one of the mirrors in the interferometer has an arbitrary deformity in the *surface figure*.[4] A new fringe stripe occurs for every half wavelength that the surface varies. (The round trip turns a half wavelength into a whole wavelength.) This makes it possible to determine the flatness of a surface with very high precision. Of course, in order to test a given surface in an interferometer, the quality of all other surfaces in the interferometer must first be ensured.

---

[3] This ignores an additive constant for fixed z.

[4] The surface figure is a name for how well a surface contour matches a desired prescription.

A typical industry standard for research-grade optics is to specify the surface flatness to within one tenth of an optical wavelength (633 nm HeNe laser). This means that the interferometer should reveal no more than one fifth of a fringe variation across the substrate surface. The fringe pattern tells the technician how the surface should continue to be polished in order to achieve the desired surface flatness. Figure 12.3(a) shows the fringe pattern for a surface with significant variations in the surface figure.

When testing a surface, it is not necessary to remove all tilt from the alignment before the effects of surface variations become apparent in the fringe pattern. In fact, it can be helpful to observe the distortions as deflections in a normally regularly striped fringe pattern. Figure 12.3(b) shows fringes from the same distorted surface when some tilt is left in the interferometer alignment. An important advantage to leaving some tilt in the beam is that one can better tell the sign of the phase errors. We can see, for example, in the case of tilt that the two major distortion regions in Fig. 12.3 have opposite phase; we can tell that one region of the substrate protrudes while the other dishes in. On the other hand, this is not clear for an interferogram with no tilt.

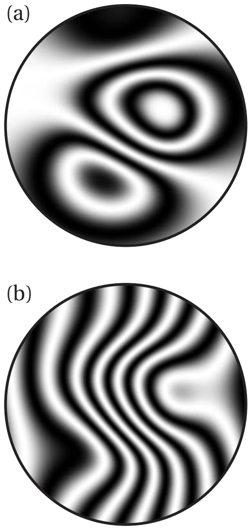


Figure 12.3 (a) Fringe pattern from an arbitrarily distorted mirror in a perfectly aligned interferometer with plane wave beams. (b) Fringe pattern from the same mirror when the mirror is tilted (still plane wave beams).

Other types of optical components (besides flat mirrors) can also be tested with an interferometer. Figure 12.4 shows how a lens can be tested using a convex mirror to compensate for the focusing action of the lens. With appropriate spacing, the lens-mirror combination can act like a flat surface. Distortions in the lens figure are revealed in the fringe pattern. In this case, the surfaces of the lens are tested together, and variations in optical path length are observed. In order to record fringes, say with a CCD camera, it is often convenient to image a larger beam onto a relatively small active area of the detector. The imaging objective should be adjusted to produce an image of the test optic on the detector screen. The diameter of the objective lens needs to accommodate the whole beam.

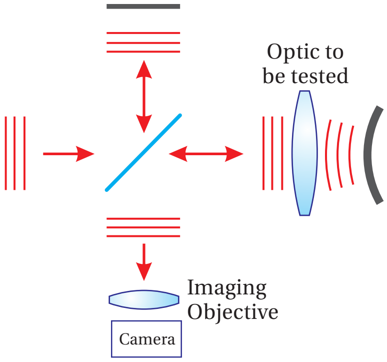


Figure 12.4 Twyman-Green setup for testing lenses.

12.3 Generating Holograms

In the late 1940's, Dennis Gabor developed the concept of holography, but it wasn't until after the invention of the laser that this field really blossomed. Consider a coherent monochromatic beam of light that is split in half by a beamsplitter, similar to that in a Michelson interferometer. Let one beam, called the reference beam, proceed directly to a recording film, and let the other beam scatter from an arbitrary object back towards the same film. The two beams interfere at the recording film. It is best to split the beam initially into unequal intensities such that the light scattered from the object has an intensity similar to the reference beam at the film.

The purpose of the film is to record the interference pattern. It is important that the coherence length of the light be much longer than the difference in path length starting from the beam splitter and ending at the film. In addition, during exposure to the film, it is important that the whole setup be stable against vibrations on the scale of a wavelength since this will cause the fringes to wash

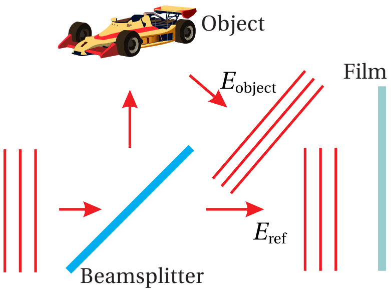
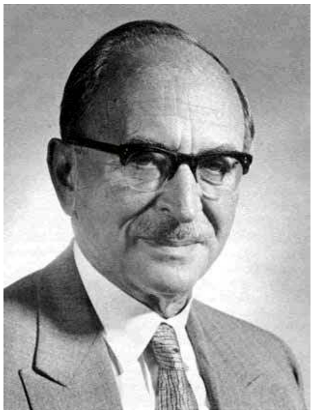


Figure 12.5 Exposure of holographic film.



Dennis Gabor (1900–1979, Hungarian) was born in Budapest. As a teenager, he fought for Hungary in World War I. Following the war, he studied at the Technical University of Budapest and later at the Technical University of Berlin. In 1927, Gabor completed his doctoral dissertation on cathode ray tubes and began a long career working on electron-beam devices such as oscilloscopes, televisions, and electron microscopes. It was in the context of 'electron optics' that he invented the concept of holography, which relied on the wave nature of electron beams. Gabor did this work while working for a British company, after fleeing Germany when Hitler came to power. Holography did not become practical until after the invention of the laser, which provided a bright coherent light source. (Gabor had attempted to make holograms earlier using a spectral line from a mercury lamp.) In 1964 the first hologram was produced. Soon after, holograms became commercially available and were popularized. Gabor accepted a post as professor of applied physics at the Imperial College of London from 1958 until he retired in 1967. He was awarded the Nobel prize in physics in 1971 for the invention of holography. (Wikipedia)

out. For simplicity, we neglect the vector nature of the electric field, assuming that the scattering from the object for the most part preserves polarization and that the angle between the two beams incident on the film is modest (so that the electric fields of the two beams are close to parallel). To the extent that the light scattered from the object contains the polarization component orthogonal to that of the reference beam, it provides a uniform (unwanted) background exposure to the film on top of which the fringe pattern is recorded.

In general terms, we may write the electric field arriving at the film as[5]

$$E_{\text{film}}(\mathbf{r})\,e^{-i\omega t} = E_{\text{object}}(\mathbf{r})\,e^{-i\omega t} + E_{\text{ref}}(\mathbf{r})\,e^{-i\omega t} \tag{12.5}$$

Here, the coordinate $\mathbf{r}$ indicates locations on the film surface, which may have arbitrary shape but often is a plane. The field $E_{\text{object}}(\mathbf{r})$, which is scattered from the object, is in general very complicated. The field $E_{\text{ref}}(\mathbf{r})$ may be equally complicated, but typically it is convenient if it has a simple form such as a plane wave, since this beam must be re-created later in order to view the hologram.

The intensity of the field (12.5) is given by

$$\begin{aligned} I_{\text{film}}(\mathbf{r}) &= \frac{1}{2}c\epsilon_0 \left|E_{\text{object}}(\mathbf{r}) + E_{\text{ref}}(\mathbf{r})\right|^2 \\ &= \frac{1}{2}c\epsilon_0 \left[\left|E_{\text{object}}(\mathbf{r})\right|^2 + \left|E_{\text{ref}}(\mathbf{r})\right|^2 + E^*_{\text{ref}}(\mathbf{r})\,E_{\text{object}}(\mathbf{r}) + E_{\text{ref}}(\mathbf{r})\,E^*_{\text{object}}(\mathbf{r})\right] \end{aligned} \tag{12.6}$$

For typical photographic film, the exposure of the film is proportional to the intensity of the light hitting it. This is known as the linear response regime. That is, after the film is developed, the transmittance T of the light through the film is proportional to the intensity of the light that exposed it (I_{film}). However, for low exposure levels, or for film specifically designed for holography, the transmission of the light through the film can be proportional to the square of the intensity of the light that exposes the film. Thus, after the film is exposed to the fringe pattern and developed, the film acquires a spatially varying transmission function according to

$$T(\mathbf{r}) \propto I^2_{\text{film}}(\mathbf{r}) \tag{12.7}$$

If at a later point in time light of intensity I_{incident} is directed onto the film, it will transmit according to $I_{\text{transmitted}} = T(\mathbf{r})\,I_{\text{incident}}$. In this case, the *field*, as it emerges from the other side of the film, will be

$$E_{\text{transmitted}}(\mathbf{r}) = t(\mathbf{r})\,E_{\text{incident}}(\mathbf{r}) \propto I_{\text{film}}(\mathbf{r})\,E_{\text{incident}}(\mathbf{r}) \tag{12.8}$$

where $t(\mathbf{r}) = \sqrt{T(\mathbf{r})}$.

12.4 Holographic Wavefront Reconstruction

To see a holographic image, we re-illuminate film (previously exposed and developed) with the original reference beam. That is, we send in

$$E_{\text{incident}}(\mathbf{r}) = E_{\text{ref}}(\mathbf{r}) \tag{12.9}$$

---

[5] See P. W. Milonni and J. H. Eberly, *Lasers*, Sect. 16.4-16.5 (New York: Wiley, 1988); G. R. Fowles, *Introduction to Modern Optics*, 2nd ed., Sect. 5.7 (Toronto: Dover, 1975).

and view the light that is transmitted. According to (12.6) and (12.8), the transmitted field is proportional to

$$
\begin{aligned}
E_{\text{transmitted}}(\mathbf{r}) &\propto I_{\text{film}}(\mathbf{r})\,E_{\text{ref}}(\mathbf{r}) \\
&= \left[\left|E_{\text{object}}(\mathbf{r})\right|^2 + \left|E_{\text{ref}}(\mathbf{r})\right|^2\right] E_{\text{ref}}(\mathbf{r}) + \left|E_{\text{ref}}(\mathbf{r})\right|^2 E_{\text{object}}(\mathbf{r}) + E_{\text{ref}}^2(\mathbf{r})\,E_{\text{object}}^*(\mathbf{r})
\end{aligned} \tag{12.10}
$$

Although (12.10) looks fairly complicated, each of the three terms has a direct interpretation. The first term is just the reference beam $E_{\text{ref}}(\mathbf{r})$ with an amplitude modified by the transmission through the film. It is the residual undeflected beam, similar to zero-order diffraction through a transmission grating. The second term is interpreted as a reconstruction of the light field originally scattered from the object $E_{\text{object}}(\mathbf{r})$. Its amplitude is modified by the intensity of the reference beam, but if the reference beam is uniform across the film, this hardly matters. An observer looking into the film sees a wavefront identical to the one produced by the original object (superimposed with the other fields in (12.10)). Thus, the observer sees a virtual image at the location of the original object. Since the wavefront of the original object has genuinely been recreated, the image looks ‘three-dimensional’, because the observer is free to view from different perspectives.

The final term in (12.10) is proportional to the complex conjugate of the original field from the object. It also contains twice the phase of the reference beam, which we can overlook if the reference beam is uniform on the film. In this case, the complex conjugate of the object field actually converges to a real image of the original object. This image is located on the observer’s side of the film, but it is often of less interest since the image is ‘inside out’. An ideal screen for viewing this real image would be an item shaped similar to the original object, which of course defeats the purpose of the hologram! To the extent that the film is not flat or to the extent that the reference beam is not a plane wave, the phase of $E_{\text{ref}}^2(\mathbf{r})$ severely distorts the image. On the other hand, the virtual image previously described never suffers from this problem.

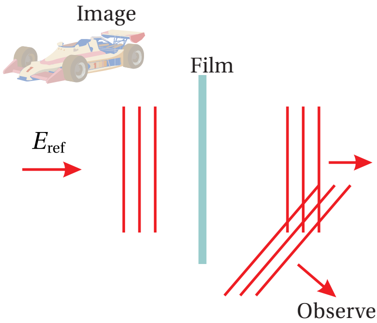


Figure 12.6 Holographic reconstruction of wavefront through diffraction from fringes on film. Compare with Fig. 12.5.

Example 12.1

Analyze the three field terms in (12.10) for a hologram made from a point object, as depicted in Fig. 12.7.

Solution: Presumably, the point object is illuminated sufficiently brightly so as to make the scattered light have an intensity similar to the reference beam at the film.

Let the reference plane wave strike the film at normal incidence. Then the reference field will have constant amplitude and phase across it; call it E_{ref}. The field from the point object can be treated as a spherical wave:

$$E_{\text{object}}(\rho) = \frac{E_{\text{ref}} L}{\sqrt{L^2+\rho^2}} e^{ik\sqrt{L^2+\rho^2}} \qquad \text{(point source example)} \tag{12.11}$$

Here ρ represents the radial distance from the center of the film to some other

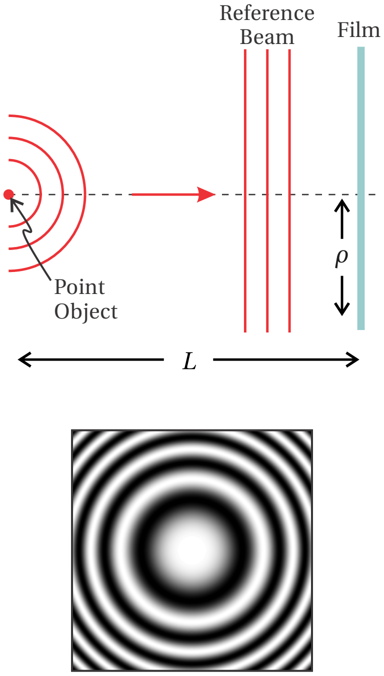


Figure 12.7 Exposure to holographic film by a point source and a reference plane wave. The holographic fringe pattern for a point object and a plane wave reference beam exposing a flat film is shown on the right.

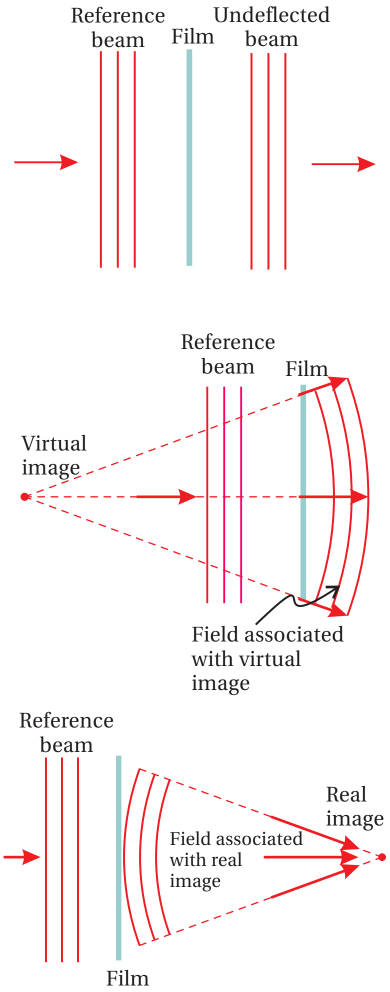

Reference beam
Film
Undeflected beam
Reference beam
Film
Virtual image
Field associated with virtual image
Reference beam
Real image
Field associated with real image
Film

Figure 12.8 Reference beam incident on previously exposed holographic film. (a) Part of the beam goes through. (b) Part of the beam takes on the field profile of the original object. undeflected. (c) Part of the beam converges to a real image of the original object.

point on the film. We have taken the amplitude of the object field to match E_{ref} in the center of the film.

After the film is exposed, developed, and re-illuminated by the reference beam, the field emerging from the right-hand-side of the film, according to (12.10), becomes

$$\begin{aligned} E_{\text{transmitted}}(\rho) \propto \left[\frac{E_{\text{ref}}^2 L^2}{L^2+\rho^2} + E_{\text{ref}}^2 \right] E_{\text{ref}} + E_{\text{ref}}^2 \frac{E_{\text{ref}} L}{\sqrt{L^2+\rho^2}} e^{ik\sqrt{L^2+\rho^2}} \\ + E_{\text{ref}}^2 \frac{E_{\text{ref}} L}{\sqrt{L^2+\rho^2}} e^{-ik\sqrt{L^2+\rho^2}} \end{aligned} \tag{12.12}$$

We see the three distinct waves that emerge from the holographic film. The first term in (12.12) represents the plane wave reference beam passing straight through the film with some variation in amplitude (depicted in Fig. 12.8 (a)). The second term in (12.12) has the identical form as the field from the original object (aside from an overall amplitude factor). It describes an outward-expanding spherical wave, which gives rise to a virtual image at the location of the original point object, as depicted in Fig. 12.8 (b). The final term in (12.12) corresponds to a converging spherical wave, which focuses to a point at a distance L from the observer's side of the screen (depicted in Fig. 12.8 (c)).

Exercises

Exercises for 12.1 Interferograms

P12.1 An ideal Michelson interferometer that uses flat mirrors is perfectly aligned to a wide collimated laser beam. Suppose that one of the mirrors is then misaligned by 0.1°. What is the spacing between adjacent fringes on the screen if the wavelength is $\lambda = 633$ nm? What would happen if, instead of tilting one of the mirrors, the angle of the input beam (before the beamsplitter) changed by 0.1°?

P12.2 An ideal Michelson interferometer uses flat mirrors perfectly aligned to an expanding beam that diverges from a point 50 cm before the beamsplitter. Suppose that one mirror is 10 cm away from the beam splitter, and the other is 11 cm. Suppose also that the center of the resulting bull's-eye fringe pattern is dark. If a screen is positioned 10 cm after the beam splitter, what is the radial distance to the next dark fringe on the screen? Take the wavelength to be $\lambda = 633$ nm.

Exercises for 12.2 Testing Optical Surfaces

L12.3 Set up an interferometer and observe distortions to a mirror substrate when the setscrew holding it is over tightened.

Exercises for 12.3 Generating Holograms

P12.4 Consider a diffraction grating as a simple hologram. Let the light from the 'object' be a plane wave (object placed at infinity) directed onto a flat film at angle θ. Let the reference beam strike the film at normal incidence, and take the wavelength to be λ.

(a) What is the period of the fringes?

(b) Show that when re-illuminated by the reference beam, the three terms in (12.10) give rise to zero-order and 1st-order diffraction (occurring on each side of zero-order).

P12.5 (a) Show that the phase of the real image in (12.12) may be approximated as $\Delta\phi = -k\rho^2/2L$, aside from a spatially independent overall phase. Compare with (11.10) and comment.

(b) This hologram is similar to a Fresnel zone plate, sometimes used to focus extreme ultraviolet light or x-rays, since it is difficult to make a lens otherwise at those wavelengths.[6] Graph the field transmission for the hologram as a function of ρ and superimpose a similar graph for a 'best-fit' mask that has regions of either 100% or 0% transmission (see Fig. 12.9). Use $\lambda = 10$ nm and $L = 10^7\lambda$ (this places the point source about 32 cm before the screen).

---

[6] Tiny Fresnel zone plates can be made for this purpose using electron-beam lithography.

L12.6 Make a hologram.

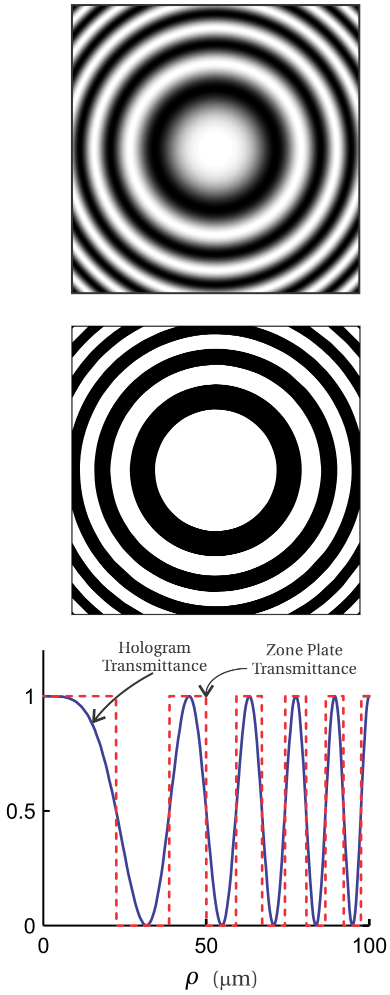


Figure 12.9 Field transmission for a point-source hologram (upper) and a Fresnel zone plate (middle), and a plot of both as a function of radius (bottom).

Review, Chapters 9–12

True and False Questions

R48 T or F: The eikonal equation and Fermat's principle depend on the assumption that the wavelength is relatively small compared to features of interest.

R49 T or F: Fermat's principle depends on the assumption that the index of refraction varies only gradually.

R50 T or F: Fermat's principle depends on the assumption that the angles involved must not be too big.

R51 T or F: The imaging relation $1/f = 1/d_o + 1/d_i$ relies on the paraxial ray approximation.

R52 T or F: Spherical aberration can be important even when the paraxial approximation works well.

R53 T or F: Chromatic aberration (the fact that refractive index depends on frequency) is an example of the violation of the paraxial approximation.

R54 T or F: The ABCD matrix for a complicated multi-element lens system can be made to look like a single thin lens through the use of principal planes.

R55 T or F: The spacing L between two flat mirrors can be chosen to make a laser cavity stable.

R56 T or F: The spherical waves given by e^{ikR}/R are exact solutions to Maxwell's equations.

R57 T or F: The Fresnel approximation falls within the paraxial approximation.

R58 T or F: Spherical waves can be used to understand diffraction from apertures that are relatively large compared to λ.

R59 T or F: The central peak of the Fraunhofer diffraction from two narrow slits separated by spacing h has the same width as the central diffraction peak from a single slit with width $\Delta x = h$.

R60 T or F: The central peak of the Fraunhofer diffraction from a circular aperture of diameter D has the same width as the central diffraction peak from a single slit with width $\Delta x = D$.

R61 T or F: The array theorem is useful for deriving *Fresnel* diffraction from a grating.

R62 T or F: A diffraction grating with a period h smaller than a wavelength is ideal for making a spectrometer.

R63 T or F: The resolving power of a spectrometer used in a particular diffraction order depends *only* on the number of lines illuminated (and not on wavelength λ or grating spacing h).

R64 T or F: The Fraunhofer diffraction pattern appearing at the focus of a lens varies in *angular* width, depending on the focal length of the lens used.

R65 T or F: Fraunhofer diffraction can be viewed as a spatial Fourier transform (or inverse transform if you prefer) on the field at the aperture.

Problems

R66 (a) Derive Snell's law using Fermat's principle.

(b) Derive the law of reflection using Fermat's principle.

R67 (a) Consider a ray of light emitted from an object, which travels a distance d_o before traversing a lens of focal length f and then traveling a distance d_i.

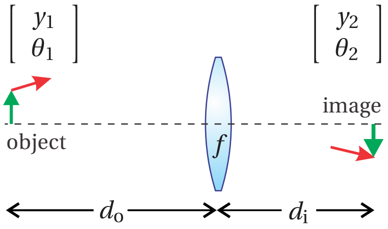

$\begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix}$
$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix}$
image
object
f
d_o
d_i

Figure 12.10

Write a vector equation relating $\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix}$ to $\begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix}$. Be sure to simplify the equation so that only one ABCD matrix is involved.

HINT: $\begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix}, \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}$

(b) Explain the requirement on the ABCD matrix in part (a) that ensures that an image appears for the distances chosen. From this requirement, extract a familiar constraint on d_o and d_i. Also, make a reasonable definition for magnification M in terms of y_1 and y_2, then substitute to find M in terms of d_o and d_i.

(c) A telescope is formed with two thin lenses separated by the sum of their focal lengths f_1 and f_2. The purpose of a telescope is to enlarge

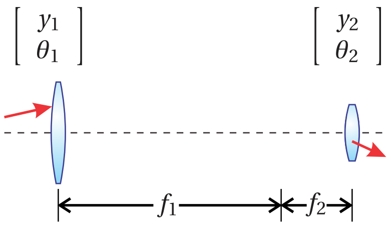


Figure 12.11

the apparent angle between points in the distant field of view. All rays entering the telescope with angle θ_1 are mapped into a (presumably) larger angle θ_2.

Give a sensible definition for angular magnification in terms of θ_1 and θ_2 and use the ABCD-matrix formulation to derive the angular magnification of the telescope in terms of f_1 and f_2.

R68 (a) Show that a system represented by a matrix $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ (beginning and ending in the same index of refraction) can be made to look like the matrix for a thin lens if suitable distances p_1 and p_2 are appended before and after the ABCD system.

HINT: $\begin{vmatrix} A & B \\ C & D \end{vmatrix} = 1$.

(b) Where are the principal planes located and what is the effective focal length for two identical thin lenses with focal lengths f that are separated by a distance $d = f$ (see Fig. 12.12)?

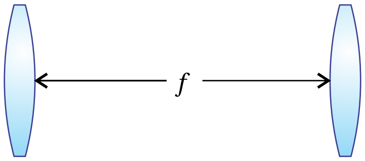


Figure 12.12

R69 Derive the *on-axis* intensity (i.e. $x, y = 0$) of a Gaussian laser beam if you know that at $z = 0$ the electric field of the beam is

$$E\left(\rho', z = 0\right) = E_0 e^{-\frac{\rho'^2}{w_0^2}}$$

Fresnel approximation:

$$E(x, y, z) \cong -\frac{i e^{ikz} e^{i\frac{k}{2z}\left(x^2+y^2\right)}}{\lambda z} \iint E\left(x', y', 0\right) e^{i\frac{k}{2z}\left(x'^2+y'^2\right)} e^{-i\frac{k}{z}\left(xx'+yy'\right)} dx' dy'$$

$$\int_{-\infty}^{\infty} e^{-Ax^2+Bx+C} dx = \sqrt{\frac{\pi}{A}} e^{\frac{B^2}{4A}+C}.$$

R70 (a) You decide to construct a simple laser cavity with a flat mirror and another mirror having concave curvature of $R = 100$ cm. What is the longest possible stable cavity that you can make?

HINT: Sylvester's theorem is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^N = \frac{1}{\sin\theta} \begin{bmatrix} A\sin N\theta - \sin(N-1)\theta & B\sin N\theta \\ C\sin N\theta & D\sin N\theta - \sin(N-1)\theta \end{bmatrix}$$

where $\cos\theta = \frac{1}{2}(A + D)$.

(b) The amplifier is YLF crystal, which lases at $\lambda = 1054$ nm. You decide to make the cavity 10 cm shorter than the longest possible (i.e. found in part (a)). What is the value of w_0, and where is the beam waist located inside the cavity (the place we assign to $z = 0$)?

HINT: For a mode to exist in a laser cavity, the radius of curvature of *each* of the end mirror matches the radius of curvature $R(z)$ of the beam at that location.

$$
\begin{aligned}
E(\rho, z) &= E_0 \frac{w_0}{w(z)} e^{-\frac{\rho^2}{w^2(z)}} e^{ikz + i\frac{k\rho^2}{2R(z)}} e^{-i\tan^{-1}\frac{z}{z_0}} \\
\rho^2 &\equiv x^2 + y^2 \\
w(z) &\equiv w_0\sqrt{1 + z^2/z_0^2} \\
R(z) &\equiv z + z_0^2/z \\
z_0 &\equiv \frac{kw_0^2}{2}
\end{aligned}
$$

R71 (a) Compute the Fraunhofer diffraction intensity pattern for a uniformly illuminated circular aperture with diameter D.

HINT: $E(x, y, z) \cong -\frac{ie^{ikz}e^{i\frac{k}{2z}(x^2+y^2)}}{\lambda z} \iint E(x', y', 0)\, e^{-i\frac{k}{z}(xx'+yy')}\, dx'dy'$

$J_0(\alpha) = \frac{1}{2\pi}\int\limits_0^{2\pi} e^{\pm i\alpha\cos(\theta-\theta')}d\theta'$, $\int\limits_0^a J_0(bx)\, x dx = \frac{a}{b}J_1(ab)$

$J_1(1.22\pi) = 0$, $\lim\limits_{x\to 0}\frac{2J_1(x)}{x} = 1$

(b) The objective lens of a telescope has a diameter $D = 30$ cm. You wish to use the telescope to examine two stars in a binary system. The stars are approximately 25 light-years away. How far apart need the stars be (in the perpendicular sense) for you to distinguish them in the visible range of $\lambda = 500$ nm? Compare with the radius of Earth's orbit, 1.5×10^8 km, often call an astronomical unit.

R72 (a) Derive the Fraunhofer diffraction pattern for the *field* from a uniformly illuminated single slit with width Δx. (Don't worry about the y-dimension.)

(b) Find the Fraunhofer *intensity* pattern for a grating with N slits of width Δx positioned on the mask at $x'_n = h\left(n - \frac{N+1}{2}\right)$ so that the spacing between all slits is h.

HINT: The array theorem says that the diffraction pattern is $\sum\limits_{n=1}^{N} e^{-i\frac{k}{z}xx'_n}$ times the diffraction pattern of a single slit. You will need

$$
\sum_{n=1}^{N} r^n = r\frac{r^N - 1}{r - 1}
$$

(c) Consider Fraunhofer diffraction from the grating in part (b). The grating is 5.0 cm wide and is uniformly illuminated. For best resolution in a monochromator with a 50 cm focal length, what should the width of the exit slit be? Assume $\lambda = 500$ nm.

R73 (a) A monochromatic plane wave with intensity I_0 and wavelength λ is incident on a circular aperture with diameter D followed by a lens with focal length f (see Fig. 12.13). What is the intensity distribution at a distance f behind the lens?

(b) You wish to 'spatially filter' the beam such that, when it emerges from the focus, it varies smoothly without diffraction rings or hard edges. A pinhole is placed at the focus, which transmits only the central portion of the Airy pattern (inside of the first zero). Calculate the intensity pattern at a distance f after the pinhole using the approximation given in the hint below.

HINT: A reasonably good approximation of the transmitted field is that of a Gaussian $E(\rho,0) = E_f e^{-\rho^2/w_0^2}$, where E_f is the magnitude of the field at the center of the focus found in part (a), and the width is $w_0 = 2\lambda f^\#/\pi$ with $f^\# \equiv f/D$. Fig. 12.14 shows how well this Gaussian approximation fits the actual curve. We have assumed that the first aperture is a distance f before the lens so that at the focus after the lens the wave front is flat. To avoid integration, you may want to use the field provided in R70 and take the far-field limit: $z >> z_0$.

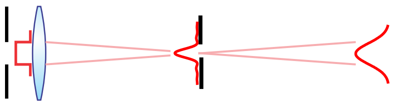


Figure 12.13 Spatial filtering of a Airy pattern.

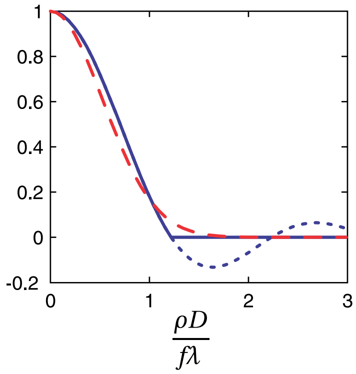


Figure 12.14 Diffraction pattern from a circular aperture (solid) that is chopped by a pinhole to remove the diffraction rings (dotted). A Gaussian field (dashed) approximates the center portion that transmits through the pinhole.

Selected Answers

R70: (a) 100 cm (b) 0.32 mm.

R71: (b) 4.8×10^8 km.

R72: (c) 5 μm.

Chapter 13

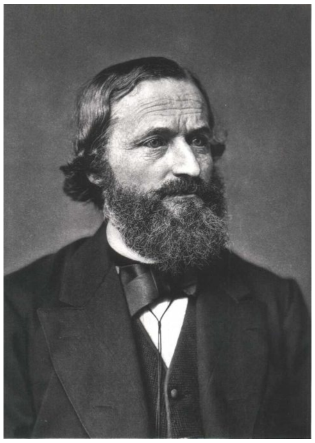
Blackbody Radiation

Hot objects glow. In 1860, Kirchhoff proposed that the radiation emitted by hot objects as a function of frequency is approximately the same for all materials.[1] The notion that all materials behave similarly led to the concept of an ideal *blackbody* radiator. Most materials have a certain shininess that causes light to reflect or scatter in addition to being absorbed and reemitted. However, light that falls upon an *ideal blackbody* is absorbed perfectly before the possibility of reemission, hence the name blackbody.

The distribution of frequencies emitted by a blackbody radiator is related to its temperature. We often consider a blackbody radiator that is in *thermal equilibrium* with the surrounding light that is absorbed and reemitted. If it is not in thermal equilibrium, for example, if more light is emitted than absorbed, then the object inevitably cools as light escapes to the environment, moving the system toward thermal equilibrium.

The Sun is a good example of a blackbody radiator. The light emitted from the Sun is associated with its surface temperature. Any light that arrives to the Sun from outer space is virtually 100% absorbed, however little light that might be, so the name *blackbody* aptly describes it. Mostly, light escapes to the much colder surrounding space (i.e. it is not in thermal equilibrium), and the temperature of the Sun's surface is maintained by the fusion process within. As another example, a glowing tungsten filament in an ordinary light bulb may be reasonably described as a blackbody radiator. However, surface reflections make it less than ideal both for absorption and emission.

Experimentally, a near perfect blackbody radiator can be constructed from a hollow object. An example is shown in Fig. 13.1. As the interior of the object is heated, the light present inside the internal cavity is in equilibrium with the glowing walls. A small hole can be drilled through the wall to observe the radiation inside without significantly disturbing the system. The observation hole can be thought of as a perfect blackbody since any light entering the hole from the outside is eventually absorbed (before being potentially reemitted), if not on the



Gustav Kirchhoff (1824–1887, German) was born in Konigsberg, the son of a lawyer. Kirchhoff attended the University of Konigsberg. While still a student, he developed what are now called Kirchhoff's law for electrical circuits. During his career, Kirchhoff was a professor in Breslau, Heidelberg, and finally Berlin. Kirchhoff was one of the first to study the spectra emitted by various objects when heated. Not coincidentally, his colleague in Heidelberg was Robert Bunsen, inventor of the Bunsen burner. Kirchhoff coined the term 'blackbody' radiation. He demonstrated that an excited gas gives off a discrete spectrum, and that an unexcited gas surrounding a blackbody emitter produces dark lines in the blackbody spectrum. Together Kirchhoff and Bunsen discovered caesium and rubidium. Later in his career, Kirchhoff showed how to derive Fresnel's diffraction formula starting from the wave equation. (Wikipedia) first bounce then on subsequent bounces inside the cavity.

---

[1] An important exception is atomic vapors, which have relatively few discrete spectral lines. However, Kirchhoff's assumption holds quite well for most solids, which are sufficiently complex.

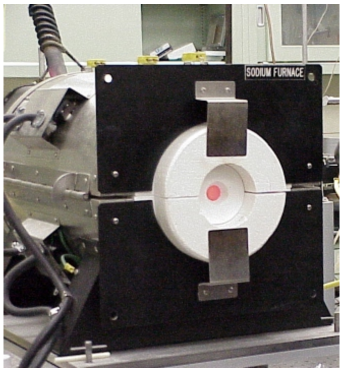


Figure 13.1 Blackbody radiator. Thermal light emerges from the small hole in the end.

In this chapter, we develop a theoretical understanding of blackbody radiation and provide some historical perspective. The explanation given by Max Planck in 1900 marks the birth of quantum mechanics. He postulated the existence of electromagnetic *quanta*, which we now call *photons*. Einstein used Planck's ideas to explain the photoelectric effect and to develop the concept of stimulated and spontaneous emission. Because of his analysis, Einstein can be thought of as the father (or maybe grandfather) of *light amplification by stimulated emission of radiation* (LASER).

13.1 Stefan-Boltzmann Law

One of the earliest properties deduced about blackbody radiation is known as the Stefan-Boltzmann law, first suggested by Stefan in 1879 and derived thermodynamically by Boltzmann in 1884.[2] This early (somewhat cumbersome) derivation is provided in appendix 13.A.[3] The Stefan-Boltzmann law says that the intensity I (including all frequencies) that flows outward from an object's surface is given by

$$I = e\sigma T^4, \tag{13.1}$$

where σ is called the Stefan-Boltzmann constant and T is the absolute temperature (in Kelvin) of the surface. The value of the Stefan-Boltzmann constant is $\sigma = 5.6696\times10^{-8}\ \mathrm{W/m^2\cdot K^4}$. The dimensionless parameter e, called the *emissivity*, is equal to one for an ideal blackbody surface. However, it takes on smaller values for actual materials because of surface reflections. For example, the emissivity of tungsten is approximately $e = 0.4$. This takes into account surface reflections, which make it harder for a material to emit light as well as to absorb light.[4]

As was mentioned in the introduction, one can construct an *ideal* blackbody radiator from a material with $e < 1$ by creating an enclosure, or *cavity*, as depicted in Fig. 13.2. A small hole in the wall behaves to the outside world like an ideal blackbody surface. From the perspective of the outside world, the hole's 'surface' has emissivity $e = 1$. Light within the cavity recirculates until it is eventually absorbed. The intensity emerging from the hole automatically approaches that of an ideal blackbody radiator.

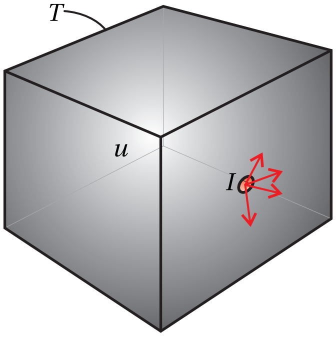


Figure 13.2 Blackbody radiator constructed as a cavity with a small hole to sample the internal light.

It is sometimes useful to express intensity in terms of the energy density of the light field u_{field} (given by (2.53) in units of energy per volume). The connection between the intensity emerging from the observation hole of a blackbody cavity and the energy density of the thermal light within the cavity is

$$I = \frac{c\,u_{\text{field}}}{4} \Rightarrow u_{\text{field}} = \frac{4\sigma T^4}{c} \tag{13.2}$$

---

[2] See P. W. Milonni, *The Quantum Vacuum An Introduction to Quantum Electrodynamics*, Sect. 1.2 (San Diego: Academic Press, 1994).

[3] It is less effort to obtain the Stefan-Boltzmann law using the Planck radiation formula as a starting point (see P13.3).

[4] Emissivity typically has some frequency dependence, so what is presented here is an oversimplification.

Within the enclosed cavity, light travels at speed c isotropically in all directions. A factor of $1/2$ arises because only half of the energy travels towards the hole from within the cavity as opposed to away. The remaining factor of $1/2$ occurs because the light emerging from the hole is directionally distributed over a hemisphere, rather than flowing in the direction of the surface normal $\hat{\mathbf{n}}$. The average over the hemisphere is carried out as follows:

$$\frac{\int_0^{2\pi} d\phi \int_0^{\pi/2} \mathbf{r}\cdot\hat{\mathbf{n}} \sin\theta\, d\theta}{\int_0^{2\pi} d\phi \int_0^{\pi/2} r \sin\theta\, d\theta} = \frac{\int_0^{2\pi} d\phi \int_0^{\pi/2} r\cos\theta \sin\theta\, d\theta}{\int_0^{2\pi} d\phi \int_0^{\pi/2} r \sin\theta\, d\theta} = \frac{1}{2} \tag{13.3}$$

Although (13.1) describes the total intensity of the light that leaves a blackbody surface, it does not describe what frequencies make up the radiation field. This frequency distribution was not fully described for another two decades, when Max Planck developed his famous formula. Planck was first to arrive at the correct formula for the spectrum of blackbody radiation, building on the work of others, most notably Wien, who came very close. At first, Planck tweaked Wien's formula to match newly available experimental data. When he attempted to explain it, he was forced to introduce the concept of light quanta. Even Planck was uncomfortable with and perhaps disbelieved the assumption that his formula implied, but he deserves credit for recognizing and articulating it.

13.2 Failure of the Equipartition Principle

In 1900, Lord Rayleigh attempted to explain the blackbody spectral distribution (intensity per frequency) as a function of temperature by applying the equipartition theorem to the problem. James Jeans gave a more complete derivation in 1905, which included an overall proportionality constant. They were hopelessly behind, since Planck nailed the answer in 1900, but their failed (classical) approach is useful pedagogically, and for that reason it gets more attention than it deserves. In this section, we also will examine the Rayleigh-Jeans approach to illustrate the shortcomings of classical concepts. This will help us better appreciate the quantum ideas in the following section. As we will see, the Rayleigh-Jeans approach actually gets the right answer in the long-wavelength limit. In fairness to Rayleigh and Jeans, they represented their formula as being useful only for long wavelengths.

The thermodynamic law of equipartition implies that the energy in a system on the average is distributed equally among all degrees of freedom in the system. For example, a system composed of oscillators (say, electrons attached to 'springs' representing the response of the material on the walls of a blackbody cavity) has an energy of $k_\mathrm{B}T/2$ for each degree of freedom, where $k_\mathrm{B} = 1.38 \times 10^{-23}\,\mathrm{J/K}$ is Boltzmann's constant. Rayleigh and Jeans supposed that each unique mode of the electromagnetic field should carry energy $k_\mathrm{B}T$ just as each mechanical spring in thermal equilibrium carries energy $k_\mathrm{B}T$ ($k_\mathrm{B}T/2$ as kinetic and $k_\mathrm{B}T/2$ as potential energy). The problem then reduces to that of finding the number of unique modes for the radiation at each frequency.[5] The idea is that requiring each mode of electromagnetic energy to hold energy $k_B T$ should reveal the spectral shape of blackbody radiation.

Number of Modes in an Electromagnetic Field

Each frequency is associated with a wave-vector magnitude $k = \sqrt{k_x^2 + k_y^2 + k_z^2}$. Notice that there are many ways (i.e. combinations of k_x, k_y, and k_z) to come up with the same $k = \omega/c$. To count these ways properly, we can let our experience with Fourier series guide us. Consider a box having length L on each side. The Fourier theorem (0.42) states that the total field inside the box (no matter how complicated the distribution) can always be represented as a superposition of sine (and cosine) waves. The total field in the box can therefore be written as[6]

$$\mathrm{Re}\left\{ \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} E_{n,m,\ell}\, e^{i(nk_0x + mk_0y + \ell k_0 z)} \right\} \tag{13.4}$$

where each component of the wave number in any of the three dimensions is an integer times

$$k_0 = 2\pi/L \tag{13.5}$$

Considering a box of size L does not artificially restrict our analysis, since we may later take the limit $L \to \infty$ so that our box represents the entire universe. Moreover, L will naturally disappear from our calculation when we later consider the *density of modes.*

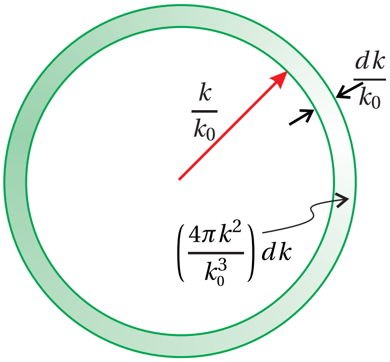


Figure 13.3 The volume of a thin spherical shell in n, m, ℓ space.

We can think of a given wave number k as specifying the equation of a sphere in a coordinate system with axes labeled n, m, and ℓ:

$$n^2 + m^2 + \ell^2 = \left(\frac{k}{k_0}\right)^2 \tag{13.6}$$

The fact that the integers n, m, and ℓ range over both positive and negative values automatically takes into account that the field may travel in the forwards or the backwards direction.

We need to know how many more ways there are to choose n, m, and ℓ when the wave number k/k_0 increases to $(k + dk)/k_0$. The answer is the difference in the volume of the two spheres shown in Fig. 13.3:

$$\#\text{ modes in }(k, k+dk) = \left(4\pi \frac{k^2}{k_0^2}\right) \frac{dk}{k_0} \tag{13.7}$$

This is the number of terms in (13.4) associated with a wave number between k and $k + dk$.

---

[5] See O. Svelto, *Principles of Lasers*, 4th ed., translated by D. C. Hanna, Sect. 2.2.1 (New York: Plenum Press, 1998).

[6] The Fourier expansion 13.4 implies that the field on the right and left of each dimension match up, which is known as periodic boundary conditions.

According to the Rayleigh-Jeans assumption, each mode should carry on average equal energy $k_\mathrm{B}T$. The energy *density* associated with a specified range of wave numbers dk is then $k_\mathrm{B}T/L^3$ times the number of modes within that range (13.7).

The total energy density in the field involving all wave numbers is then[7]

$$u_\text{field} = \int_0^\infty 2 \times \frac{k_\mathrm{B}T}{L^3} \times \frac{4\pi k^2}{k_0^3} dk = k_\mathrm{B}T \int_0^\infty \frac{k^2}{\pi^2} dk \tag{13.8}$$

where the extra factor of 2 accounts for two independent polarizations, not specified in (13.4). As anticipated, the dependence on L has disappeared from (13.8) after substituting from (13.5).

We can immediately see that (13.8) disagrees drastically with the Stefan-Boltzmann law (13.2), since (13.8) is proportional to temperature rather than to its fourth power. In addition, the integral in (13.8) is seen to diverge, meaning that regardless of the temperature, the light carries infinite energy density! This has since been named the *ultraviolet catastrophe* since the divergence occurs on the short wavelength end of the spectrum. This is a clear failure of classical physics to explain blackbody radiation. Nevertheless, Rayleigh emphasized the fact that his formula works well for the longer wavelengths.

It is instructive to make the change of variables $k = \omega/c$ in the integral to write

$$u_\text{field} = k_\mathrm{B}T \int_0^\infty \frac{\omega^2}{\pi^2 c^3} d\omega \tag{13.9}$$

The important factor $\omega^2/\pi^2 c^3$ can now be understood to be the number of modes per frequency. Then (13.9) is rewritten as

$$u_\text{field} = \int_0^\infty \rho(\omega)\, d\omega \tag{13.10}$$

where

$$\rho_\text{Rayleigh-Jeans}(\omega) = k_\mathrm{B}T \frac{\omega^2}{\pi^2 c^3} \tag{13.11}$$

describes (incorrectly) the *spectral energy density* of the radiation field associated with blackbody radiation.

James Jeans (1877–1946, English) was born in Ormskirk, England. He attended Cambridge University and later taught there for most of his career. He also taught at Princeton University for a number of years. One of his major contributions was the development of Jeans length, the critical radius for interstellar clouds, which determines whether a cloud will collapse to form a star. In his later career, Jeans became somewhat well known to the public for his lay-audience books highlighting scientific advances, in particular relativity and cosmology. (Wikipedia)

13.3 Planck's Formula

In the late 1800's as spectrographic technology improved, experimenters acquired considerable data on the spectra of blackbody radiation. For the first time, detailed maps of the intensity per frequency associated with blackbody radiation became available over a fairly wide wavelength range. In keeping with Kirchhoff's notion of an ideal blackbody radiator, the results were observed to be independent of the material for most solids. The intensity per frequency depended only on temperature and when integrated over all frequencies agreed with the Stefan-Boltzmann law (13.1).

---

[7] See O. Svelto, *Principles of Lasers*, 4th ed., translated by D. C. Hanna, Sect. 2.2.2 (New York: Plenum Press, 1998).

In 1896, Wilhelm Wien considered the known physical and mathematical constraints on the spectrum of blackbody radiation and proposed a spectral function that seemed to work:[8]

$$\rho_{\text{Wien}}(\omega) = \frac{\hbar\omega^3 e^{-\hbar\omega/k_{\text{B}}T}}{\pi^2 c^3} \tag{13.12}$$

An important feature of (13.12) is that it gives a result proportional to T^4 when integrated over all frequency ω (i.e. the Stefan-Boltzmann law).

Wien's formula did a fairly good job of fitting the experimental data. However, in 1900 Lummer and Pringshein, colleagues of Max Planck, reported experimental data that deviated from the Wien distribution at long wavelengths (infrared). Planck was privy to this information early on and introduced a modest revision to Wien's formula that fit the data beautifully everywhere:

$$\rho_{\text{Planck}}(\omega) = \frac{\hbar\omega^3}{\pi^2 c^3\left[e^{\hbar\omega/k_{\text{B}}T} - 1\right]} \tag{13.13}$$

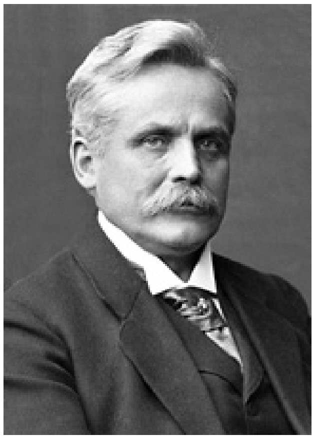
where $\hbar = 1.054 \times 10^{-34}\,\text{J}\cdot\text{s}$ is an experimentally determined constant.[9]

Figure 13.4 shows the Planck spectral distribution curve together with the Rayleigh-Jeans curve (13.11) and the Wien curve (13.12). As is apparent, the Wien distribution does a good job nearly everywhere. However, at long wavelengths it was off by just enough for the experimentalists to notice that something was wrong.

At this point, it may seem fair to ask, "What did Planck do that was so great?" After all, he simply guessed a function that was only a slight modification of Wien's distribution. And he knew the 'answer from the back of the book', namely Lummer's and Pringshein's well done experimental results. (At the time, Planck was unaware of the work by Rayleigh.)

Planck gets well-deserved credit for interpreting the meaning of his new formula. His interpretation was what he called an "act of desperation." He did not necessarily believe in the implications of his formula; in fact, he presented them somewhat apologetically. It was several years later that the young Einstein published his paper explaining the photoelectric effect in the context of Planck's work.

Planck's insight was an enormous step toward understanding the quantum nature of light. Nevertheless, it took another three decades to develop a more



Wilhelm Wien (1864–1928, German) was born in Gaffken, Prussia (now Primorsk, Russia). As a teenager, he attended schools in Rastenburg and then Heidelberg. He later attended the University of Göttingen and then the University of Berlin. In 1886, he received his Ph.D. after working under Hermann von Helmholtz where he studied the influence of materials on the color of light. In 1896 Wien developed an empirical formula for the spectral distribution of blackbody radiation. He collaborated with Planck, who gave the law a foundation in electromagnetic and thermodynamic theory. Planck later improved the formula, whereupon it became known by his name. However, Wien's formula for the peak wavelength of the blackbody curve, called Wien's displacement law, remains valid. In 1898, Wien identified a positive particle equal in mass to the hydrogen atom, which was later named the proton. Wien received the Nobel prize in 1911 for his work on heat and radiation. (Wikipedia)

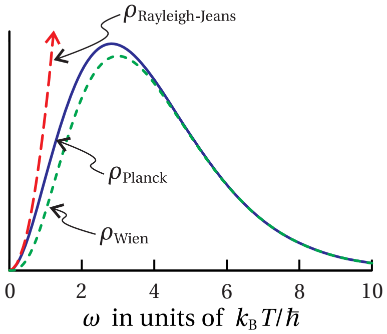


Figure 13.4 Energy density per frequency according to Planck, Wien, and Rayleigh-Jeans.

---

[8] The constant h had not yet been introduced by Planck. The actual way that Wien wrote his distribution was $\rho_{\text{Wien}}(\omega) = a\omega^3 e^{-b\omega/T}$, where a and b were parameters used to fit the data.

[9] Planck's constant was first introduced as $h = 6.626 \times 10^{-34}\,\text{J}\cdot\text{s}$, convenient for working with frequency ν, expressed in Hz. It is common to write $\hbar \equiv h/2\pi$ when working with frequency ω, expressed in rad/s.

complete theory of quantum electrodynamics. Students can take comfort in the fact that the very people who developed quantum mechanics were also bothered by its confrontation with deep-seated intuition. If quantum mechanics bothers you, you are in good company!

Max Planck (1858–1947, German) was born in Kiel, the sixth child in his family. His father was a law professor. When Max was about nine years old, his family moved to Munich where he attended gymnasium. A mathematician, Herman Muller took an interest in his schooling and tutored him in mechanics and astronomy. Planck was a gifted musician, but he decided to pursue a career in physics. At age 16 he enrolled in the University of Munich. By age 22, he had finished his doctoral dissertation and habilitation thesis. He was initially ignored by the academic community and worked for a time as an unpaid lecturer. He became an associate professor of theoretical physics at the University of Kiel and then a few years later took over Kirchhoff's post at the University of Berlin. After nearly twenty years of idyllic and happy family life, a series of tragedies hit the Planck household. Planck's first wife and mother of four, died. Then his eldest son was killed in action during World War I. Soon after, his twin daughters each died giving birth to their first child. Later Planck's remaining son from his first marriage was executed for participating in a failed attempt to assassinate Hitler. Planck won the Nobel prize in 1918 for his introduction of energy quanta, but he had serious reservations about the course that quantum mechanics theory took. (Wikipedia)

Planck found that he could derive his formula only if he made the following strange assumption: A given mode of the electromagnetic field is not able to carry an arbitrary amount of energy (for example, $k_B T$ as Rayleigh and Jeans used, which varies continuously as a function of temperature). Rather, a field mode can only carry discrete amounts of energy separated by spacing $\hbar\omega$. Under this assumption, the probability P_n that a mode of the field is excited to the n^{th} level is proportional to the Boltzmann statistical weighting factor $e^{-n\hbar\omega/k_B T}$. A review of the Boltzmann factor is given in Appendix 13.B.

Probable Energy in Each Field Mode

The Boltzmann factor can be normalized by dividing by the sum of all such factors to obtain the probability of having energy $n\hbar\omega$ in a particular mode:

$$P_n = \frac{e^{-n\hbar\omega/k_B T}}{\sum\limits_{m=0}^{\infty} e^{-m\hbar\omega/k_B T}} = e^{-n\hbar\omega/k_B T}\left[1 - e^{-\hbar\omega/k_B T}\right] \tag{13.14}$$

We used (0.66) to accomplish the above sum, which is a geometric series.

The expected energy in a particular mode of the field is the sum of each possible energy level (i.e. $n\hbar\omega$) times the probability of its occurrence:

$$\begin{aligned}
\sum_{n=0}^{\infty} n\hbar\omega P_n &= \hbar\omega\left[1 - e^{-\hbar\omega/k_B T}\right]\sum_{n=0}^{\infty} n e^{-n\hbar\omega/k_B T} \\
&= \hbar\omega\left[1 - e^{-\hbar\omega/k_B T}\right]\frac{\partial}{\partial\left(\hbar\omega/k_B T\right)}\sum_{n=0}^{\infty} e^{-n\hbar\omega/k_B T} \\
&= -\hbar\omega\left[1 - e^{-\hbar\omega/1-k_B T}\right]\frac{\partial}{\partial\left(\hbar\omega/k_B T\right)}\frac{1}{1 - e^{-\hbar\omega/k_B T}} \\
&= \frac{\hbar\omega}{e^{\hbar\omega/k_B T} - 1}
\end{aligned} \tag{13.15}$$

We used (0.66) again as well as a clever derivative trick.

Equation (13.15) provides the expected energy in any of the modes of the radiation field, as dictated by Planck's assumption. To obtain the Planck distribution (13.13), we replace $k_B T$ in the Rayleigh-Jeans formula (13.10) with the correct expected energy (13.15).[10]

It is interesting that we are now able to *derive* the constant in the Stefan-Boltzmann law (13.2) in terms of Planck's constant $\hbar$ (see P13.3). The Stefan-Boltzmann law is obtained by integrating the spectral density function (13.13) over all frequencies to obtain the total field energy density, which is in thermal equilibrium with the blackbody radiator:

---

[10] See O. Svelto, *Principles of Lasers*, 4th ed., translated by D. C. Hanna, Sect. 2.2.2 (New York: Plenum Press, 1998).

$$u_{\text{field}} = \int_0^\infty \rho_{\text{Planck}}(\omega)\, d\omega = \frac{4}{c}\frac{\pi^2 k_B^4}{60 c^2 \hbar^3} T^4 \equiv \frac{4}{c}\sigma T^4 \tag{13.16}$$

Since Planck's constant was not introduced until a couple decades after the Stefan-Boltzmann law was developed, one might more appropriately say that the Stefan-Boltzmann constant pins down Planck's constant.

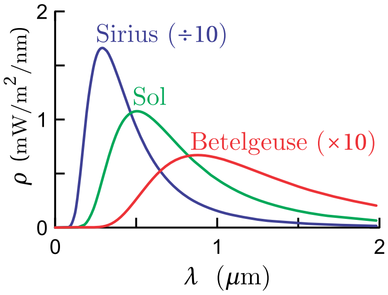


Figure 13.5 Blackbody spectrum (13.17) plotted for the surface temperature of three stars: Sirius (9900 K), the Sun (5750 K), and Betelgeuse (3300 K).

Example 13.1

Determine $\rho_{\text{Planck}}(\lambda)$ such that

$$u_{\text{field}} = \int_0^\infty \rho_{\text{Planck}}(\omega)\, d\omega = \int_0^\infty \rho_{\text{Planck}}(\lambda)\, d\lambda$$

where $\rho_{\text{Planck}}(\omega)$ and $\rho_{\text{Planck}}(\lambda)$ represent distinct functions distinguished by their arguments.

Solution: The change of variables $\lambda \equiv 2\pi c/\omega \Rightarrow d\omega = -2\pi c\, d\lambda/\lambda^2$ gives

$$u_{\text{field}} = \int_\infty^0 \frac{\hbar\,(2\pi c/\lambda)^3}{\pi^2 c^3 \left[e^{\hbar(2\pi c/\lambda)/k_B T} - 1\right]} \left(-2\pi c \frac{d\lambda}{\lambda^2}\right) = \int_0^\infty \frac{16\hbar c}{\lambda^5 \left[e^{2\pi\hbar c/\lambda k_B T} - 1\right]} d\lambda$$

By inspection, we get

$$\rho_{\text{Planck}}(\lambda) = \frac{8\pi h c}{\lambda^5 \left[e^{hc/\lambda k_B T} - 1\right]} \tag{13.17}$$

where we have written $h \equiv 2\pi\hbar$. It is interesting to note that the maximum of $\rho_{\text{Planck}}(\lambda)$ occurring at $\lambda_{\max}$ and the maximum of $\rho_{\text{Planck}}(\omega)$, occurring at $\omega_{\max}$, do not correspond to a matching wavelength and frequency. That is, $\lambda_{\max} \neq 2\pi c/\omega_{\max}$, because of the nonlinear nature of the variable transformation. (See problem P13.4.)

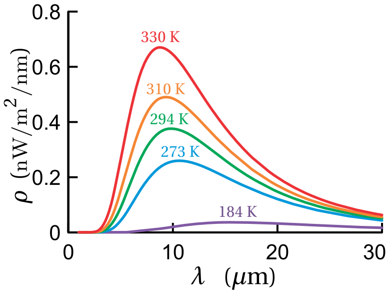


Figure 13.6 Blackbody spectrum (13.17) plotted for Earth's record cold temperature (184 K Antarctica), a typical winter day (273 K), room temperature (294 K), a typical summer day (310 K), Earth's record hot temperature (330 K Death Valley).

13.4 Einstein's A and B Coefficients

More than a decade after Planck introduced his formula, and after Niels Bohr had proposed that electrons occupy discrete energy states in atoms, Einstein reexamined blackbody radiation in terms of Bohr's new idea. If the material of a blackbody radiator interacts with a particular mode of the field with frequency ω, then electrons in the material must make transitions between two energy levels with energy separation $\hbar\omega$. Since the radiation of a blackbody is in thermal equilibrium with the material, Einstein postulated that the field *stimulates*

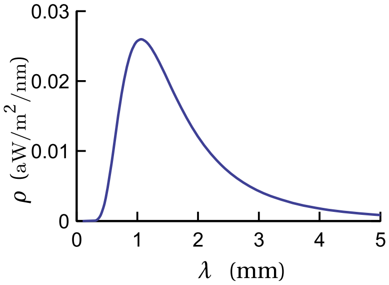


Figure 13.7 Blackbody spectrum (13.17) of the cosmic microwave background radiation that pervades the universe (2.7 K).

electron transitions between energy levels. In addition, he postulated that some transitions must occur spontaneously.

Einstein wrote down rate equations for populations of the two levels N_1 and N_2 associated with the transition $\hbar\omega$:[11]

$$\begin{aligned}\dot{N}_1 &= A_{21}N_2 - B_{12}\rho(\omega)N_1 + B_{21}\rho(\omega)N_2,\\ \dot{N}_2 &= -A_{21}N_2 + B_{12}\rho(\omega)N_1 - B_{21}\rho(\omega)N_2\end{aligned} \tag{13.18}$$

The coefficient A_{21} is the rate of spontaneous emission from state 2 to state 1, $B_{12}\rho(\omega)$ is the rate of stimulated absorption from state 1 to state 2, and $B_{21}\rho(\omega)$ is the rate of stimulated emission from state 2 to state 1. He supposed that the rate of stimulated transitions ought to be proportional to spectral density $\rho(\omega)$.

In thermal equilibrium, the rate equations (13.18) are both equal to zero (i.e., $\dot{N}_1 = \dot{N}_2 = 0$), since the relative populations of each level must remain constant. We can then solve for the spectral density $\rho(\omega)$ at the given frequency. In this case, either expression in (13.18) yields

$$\rho(\omega) = \frac{A_{21}}{\frac{N_1}{N_2}B_{12} - B_{21}} \tag{13.19}$$

In thermal equilibrium, the spectral density must match the Planck spectral density formula (13.13). In making the comparison, we should first rewrite the ratio N_1/N_2 of the populations in the two levels using the Boltzmann probability factor (see appendix 13.B):

$$\frac{N_1}{N_2} = \frac{e^{-E_1/k_B T}}{e^{-E_2/k_B T}} = e^{(E_2-E_1)/k_B T} = e^{\hbar\omega/k_B T} \tag{13.20}$$

Then when equating (13.19) to the Planck blackbody spectral density (13.13) we get

$$\frac{A_{21}}{e^{\hbar\omega/k_B T}B_{12} - B_{21}} = \frac{\hbar\omega^3}{\pi^2 c^3\left[e^{\hbar\omega/k_B T} - 1\right]} \tag{13.21}$$

From this expression we deduce that[12]

$$B_{12} = B_{21} \tag{13.22}$$

and

$$A_{21} = \frac{\hbar\omega^3}{\pi^2 c^3}B_{21} \tag{13.23}$$

We see from (13.22) that the rate of stimulated absorption is the same as the rate of stimulated emission. In addition, if one knows the rate of stimulated emission between a pair of states, it follows from (13.23) that one also knows the rate of

Albert Einstein (1879–1955, German) is without a doubt the most famous scientist in history. Time Magazine named him Person of the Century. Born in Ulm to a (non-practicing) Jewish family, young Albert was influenced by a medical student, Max Talmud, who took meals with his family and enthusiastically introduced the 10-year-old Albert to geometry and other topics. Einstein's father wanted Albert to be trained as an electrical engineer, but Albert clashed with his teachers in that program and withdrew. Einstein then attended school in Switzerland, and subsequently entered a mathematics program at the Polytechnic in Zurich. There, Einstein met his first wife, Mileva Maric, a fellow math student, who he later divorced before marrying Elsa Lowenthal. Early on, Einstein could not find a job as a professor, and so he worked in the Swiss patent office until his "Miracle Year" (1905), when he published four major papers, including relativity and the photoelectric effect (for which he later received the Nobel prize). Thereafter, job offers were never in short supply. In 1933, as the Nazi regime came to power, Einstein emigrated from Germany to the US and became a professor at Princeton University. Einstein is most noted for special and general relativity, for which he became a celebrity scientist in his own lifetime. Einstein also made huge contributions to statistical and quantum mechanics. (Wikipedia) spontaneous emission. This is remarkable because to derive A_{21} directly, one needs quantum electrodynamics (the complete photon description). However, to obtain B_{21}, it is actually only necessary to use a *semiclassical* theory, where the light is treated classically and the energy levels in the material are treated quantum-mechanically using the Schrödinger equation.

---

[11] See P. W. Milonni, *The Quantum Vacuum An Introduction to Quantum Electrodynamics*, Sect. 1.8 (San Diego: Academic Press, 1994).

[12] We assume that energy levels 1 and 2 are non-degenerate. Some modifications must be made in the case of degenerate levels, but the procedure is similar.

In writing the rate equations, (13.18), Einstein predicted the possibility of creating lasers fifty years in advance of their development. These rate equations are still valid even if the light is not in thermal equilibrium with the material. The equations suggest that if the population in the upper state 2 can be made artificially large, then amplification will result via the stimulated transition. The rate equations also show that a population inversion (more population in the upper state than in the lower one) cannot be achieved by 'pumping' the material with the same frequency of light that one hopes to amplify. This is because the stimulated absorption rate is balanced by the stimulated emission rate. The material-dependent parameters A_{21} and $B_{12} = B_{21}$ are called the Einstein A and B coefficients.

Appendix 13.A Thermodynamic Derivation of the Stefan-Boltzmann Law

In this appendix, we derive the Stefan-Boltzmann law without relying on the Planck blackbody formula.[13] This derivation is included mainly for historical interest. The derivation relies on the 1st and 2nd laws of thermodynamics.

Consider a container whose walls are all at the same temperature and in thermal equilibrium with the radiation field inside. Notice that the units of energy density u_{field} (energy per volume) are equivalent to force per area, or in other words pressure. It turns out that the radiation exerts a pressure of

$$P = u_{\text{field}}/3 \tag{13.24}$$

on the walls of the container. This can be derived from the fact that radiation of energy ΔE imparts a momentum

$$\Delta p = \frac{\Delta E}{c}\cos\theta \tag{13.25}$$

when it is absorbed with incident angle θ on a surface.[14] A similar momentum is imparted when radiation is emitted.

Derivation of (13.24)

---

[13] See P. W. Milonni, *The Quantum Vacuum An Introduction to Quantum Electrodynamics*, Sect. 1.2 (San Diego: Academic Press, 1994).

[14] The fact that light carries momentum was understood well before the development of the theory of relativity and the photon description of light.

Consider a thin layer of space adjacent to a container wall with area A. If the layer has thickness Δz, then the volume in the layer is $A\Delta z$. Half of the radiation inside the layer flows toward the wall, where it is absorbed. The total energy in the layer that will be absorbed is then $\Delta E = (A\Delta z)u_{\text{field}}/2$, which arrives during the interval $\Delta t = \Delta z/(c\cos\theta)$, assuming for the moment that all light is directed with angle θ; we must average the angle of light propagation over a hemisphere.

The pressure on the wall due to absorption (i.e. force or dp/dt per area) is then

$$P_{\text{abs}} = \frac{\int_0^{2\pi} d\phi \int_0^{\pi/2} \frac{\Delta p}{\Delta t}\frac{1}{A}\sin\theta\, d\theta}{\int_0^{2\pi} d\phi \int_0^{\pi/2} \sin\theta\, d\theta} = \frac{u_{\text{field}}}{2}\int_0^{\pi/2} \cos^2\theta \sin\theta\, d\theta = \frac{u_{\text{field}}}{6} \tag{13.26}$$

In equilibrium, an equal amount of radiation is also emitted from the wall. This gives an additional pressure $P_{\text{emit}} = P_{\text{abs}}$, which confirms that the total pressure is given by (13.24).

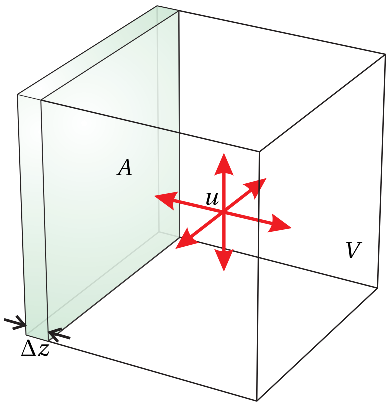


Figure 13.8 Field inside a blackbody radiator.

We derive the Stefan-Boltzmann law using the concept of entropy, which is defined in differential form by the quantity

$$dS \equiv \frac{dQ}{T} \tag{13.27}$$

where dQ is the injection of heat (or energy) into the radiation field in the box and T is the temperature at which that injection takes place. We would like to write dQ in terms of u_{field}, V, and T. Then we may invoke the fact that S is a state variable, which implies

$$\frac{\partial^2 S}{\partial T \partial V} = \frac{\partial^2 S}{\partial V \partial T} \tag{13.28}$$

This is a mathematical statement of the fact that S is fully defined if the internal energy, temperature, and volume of a system are specified. That is, S does not depend on past temperature and volume history; it is dictated by the present state of the system.

To obtain dQ in the form that we need, we can use the 1st law of thermodynamics. It states that a change in internal energy $dU = d(u_{\text{field}}V)$ can take place by the injection of heat dQ or by doing work $dW = PdV$ as the volume increases:

$$\begin{aligned} dQ &= dU + PdV = d(u_{\text{field}}V) + PdV \\ &= V\,du_{\text{field}} + u_{\text{field}}dV + \frac{1}{3}u_{\text{field}}dV \\ &= V\frac{du_{\text{field}}}{dT}dT + \frac{4}{3}u_{\text{field}}dV \end{aligned} \tag{13.29}$$

We have used energy density times volume to obtain the total energy U in the radiation field in the box. We have also used (13.24) to obtain the work accomplished by pressure as the volume changes.

We can use (13.29) to rewrite (13.27) as

$$dS = \frac{V}{T}\frac{du_{\text{field}}}{dT}dT + \frac{4u_{\text{field}}}{3T}dV \tag{13.30}$$

When we differentiate (13.30) with respect to temperature or volume we get

$$\begin{aligned}\frac{\partial S}{\partial V} &= \frac{4u_{\text{field}}}{3T}\\ \frac{\partial S}{\partial T} &= \frac{V}{T}\frac{du_{\text{field}}}{dT}\end{aligned} \tag{13.31}$$

We are now able to evaluate the partial derivatives in (13.28), which give

$$\begin{aligned}\frac{\partial^2 S}{\partial T\partial V} &= \frac{4}{3}\frac{\partial}{\partial T}\frac{u_{\text{field}}}{T} = \frac{4}{3}\frac{1}{T}\frac{\partial u_{\text{field}}}{\partial T} - \frac{4}{3}\frac{u_{\text{field}}}{T^2}\\ \frac{\partial^2 S}{\partial V\partial T} &= \frac{1}{T}\frac{du_{\text{field}}}{dT}\end{aligned} \tag{13.32}$$

Since by (13.28) these two expressions must be equal, we get a differential equation relating the internal energy of the system to the temperature:

$$\frac{4}{3}\frac{1}{T}\frac{\partial u_{\text{field}}}{\partial T} - \frac{4}{3}\frac{u_{\text{field}}}{T^2} = \frac{1}{T}\frac{du_{\text{field}}}{dT} \Rightarrow \frac{\partial u_{\text{field}}}{\partial T} = \frac{4u_{\text{field}}}{T} \tag{13.33}$$

The solution to this differential equation is (13.2), where $4\sigma/c$ is a constant to be determined experimentally.

Appendix 13.B Boltzmann Factor

The entropy of an object is defined by

$$S_{\text{obj}} = k_{\text{B}} \ln n_{\text{obj}} \tag{13.34}$$

which depends on the number of configurations n_{obj} for a given state (defined, for example, by fixed energy and volume). Now imagine that the object is placed in contact with a very large thermal reservoir. The ‘object’ could be the electromagnetic radiation inside a hollow blackbody apparatus, and the reservoir could be the walls of the apparatus, capable of holding far more energy than the light field can hold. The condition for thermal equilibrium between the object and the reservoir is

$$\frac{\partial S_{\text{obj}}}{\partial U_{\text{obj}}} = \frac{\partial S_{\text{res}}}{\partial U_{\text{res}}} \equiv \frac{1}{T} \tag{13.35}$$

where temperature has been introduced as a definition, which is consistent with (13.27).

The total number of configurations for the combined system is $N = n_{\text{obj}}n_{\text{res}}$, where n_{obj} and n_{res} are the number of configurations available within the object and the reservoir separately. A thermodynamic principle is that all possible configurations are equally probable. In thermal equilibrium, the probability for a *given* configuration in the object is therefore proportional to

$$P \propto \frac{N}{n_{\mathrm{obj}}} = n_{\mathrm{res}} = e^{S_{\mathrm{res}}/k_{\mathrm{B}}} \tag{13.36}$$

where we have invoked (13.34).

Meanwhile, a Taylor's series expansion of S_{res} yields

$$S_{\mathrm{res}}\left(U_{\mathrm{res}}\right) \cong S_{\mathrm{res}}\left(U_{\mathrm{res}}^{\mathrm{eq}}\right) + \left.\frac{\partial S_{\mathrm{res}}}{\partial U_{\mathrm{res}}}\right|_{U_{\mathrm{res}}^{\mathrm{eq}}} \left(U_{\mathrm{res}} - U_{\mathrm{res}}^{\mathrm{eq}}\right) + \ldots \tag{13.37}$$

Higher order terms are not needed since we assume the reservoir to be very large so that it is disturbed only slightly by variations in the object. Since the overall energy of the system is fixed, we may write

$$U_{\mathrm{res}} - U_{\mathrm{res}}^{\mathrm{eq}} = \Delta U_{\mathrm{res}} = -\Delta U_{\mathrm{obj}} \tag{13.38}$$

where ΔU_{obj} is a small change in energy in the object. When (13.35), (13.37), and (13.38) are introduced into (13.36), the probability for the specific configuration becomes $P \propto e^{\frac{1}{k_{\mathrm{B}}} S_{\mathrm{res}}\left(U_{\mathrm{res}}^{\mathrm{eq}}\right) - \frac{\Delta U_{obj}}{k_{\mathrm{B}} T}}$, or simply

$$P \propto e^{-\frac{\Delta U_{obj}}{k_{\mathrm{B}} T}} \tag{13.39}$$

since the first term in the exponent is constant. ΔU_{obj} represents an amount energy added to the object to establish a configuration. In the case of blackbody radiation, a mode takes on energy $\Delta U_{\mathrm{obj}} = n\hbar\omega$, where n is the number of energy quanta in the mode. The probability that a mode carries energy $n\hbar\omega$ is therefore proportional to $e^{-\frac{n\hbar\omega}{k_{\mathrm{B}} T}}$.

Exercises

Exercises for 13.1 Stefan-Boltzmann Law

P13.1 The Sun has a radius of $R_S = 6.96 \times 10^8$ m. What is the total power that it radiates, given a surface temperature of 5750 K?

P13.2 A 1 cm-radius spherical ball of polished gold hangs suspended inside an evacuated chamber that is at room temperature 20°C. There is no pathway for thermal conduction to the chamber wall.

(a) If the gold is at a temperature of 100°C, what is the *initial* rate of temperature loss in °C/s? The emissivity for polished gold is $e = 0.02$. The specific heat of gold is 129 J/kg·°C and its density is 19.3 $\mathrm{g/cm^3}$.

HINT: $Q = mc\Delta T$ and Power $= Q/\Delta t$. You should consider the power flowing both ways.

(b) What is the *initial* rate of temperature loss if the ball is coated with flat black paint, which has emissivity $e = 0.95$?

Exercises for 13.3 Planck's Formula

P13.3 Derive (or try to derive) the Stefan-Boltzmann law by integrating the

(a) Rayleigh-Jeans energy density

$$u_{\text{field}} = \int_0^\infty \rho_{\text{Rayleigh-Jeans}}(\omega)\, d\omega$$

Please comment.

(b) Wien energy density

$$u_{\text{field}} = \int_0^\infty \rho_{\text{Wien}}(\omega)\, d\omega$$

Please evaluate σ.

HINT: $\int_0^\infty x^3 e^{-ax} dx = \frac{6}{a^4}$.

(c) Planck energy density

$$u_{\text{field}} = \int_0^\infty \rho_{\text{Planck}}(\omega)\, d\omega$$

Please evaluate σ. Compare results of (b) and (c).

HINT: $\int_0^\infty \frac{x^3 dx}{e^{ax}-1} = \frac{\pi^4}{15a^4}$.

P13.4 (a) Derive Wien's displacement law

$$\lambda_{\max} = \frac{0.00290\ \mathrm{m \cdot K}}{T}$$

which gives the strongest wavelength present in the blackbody spectral distribution.

HINT: See Example 13.1. You may like to know that the solution to the transcendental equation $(5 - x)\, e^{x} = 5$ is $x = 4.965$.

(b) What is the strongest wavelength emitted by the Sun, which has a surface temperature of 5750 K (see P13.1)?

(c) Repeat the problem to find $\omega_{\max}$ and show that $\lambda_{\max} \frac{\omega_{\max}}{2\pi} \neq c$. (We naturally observe $\lambda_{\max}$ when making a measurement using a grating spectrometer.) HINT: The solution to $(3 - x)\, e^{x} = 3$ is $x = 2.821$

Index

ABCD law for Gaussian beams, 294
ABCD matrices, 236, 240
ABCD matrix, 238
aberrations, 237, 250
absolute value, complex number, 10
absolute value, vector, 1
Airy pattern, 283
Ampere's law, 25, 30
Ampere-Maxwell, 31
angle-addition formula, 7
anisotropic medium, 121
aperture, 260
Arago, Francois Jean Dominique, 263
array theorem, 277, 284
arrival time, 183
astigmatism, 251
autocorrelation theorem, 24

Babinet's principle, 262
beam waist, 277, 290, 293
Bessel function, 19, 268, 282
biaxial crystal, 127
Biot, Jean-Baptiste, 28
Biot-Savart law, 28
birefringence, 121, 127, 130
blackbody radiation, 319
blaze angle, 300
Bohr, Niels, 326
Boltzmann factor, 330
boundary conditions at an interface, 75, 84
Brewster's angle, 80
Brewster, David, 80
broadband, 169

carrier frequency, 180
Cartesian coordinates, 1
causality, 189, 193, 194
centroid, 191
characteristic matrix, 109
chirp, 181, 183
chirped pulse amplification, 187
chromatic aberration, 251
circular polarization, 143, 144
circular polarizer, 158
Clausius-Mossotti, 51, 62
coefficient of finesse, 95, 99
coherence length, 207, 208
coherence time, 207, 208
color, 58
color matching function, 61
color space, 60
coma, 252
complex angle, 11
complex conjugate, 10
complex notation, 43, 45
complex numbers, 6
complex plane, 10
complex polar representation, 9
concave, 239
conductivity, 69
conductor, refractive index of, 52
constitutive relation, 47
constitutive relation in crystals, 121
continuity equation, 31
continuous source, temporal coherence, 209
convex, 239
convolution theorem, 23
cosine, complex representation, 7
Coulomb's law, 26
critical angle, 81
cross product, 2

crystal, Poynting vector in, 130
crystal, wave propagation in, 123
curl, 3
current density, 27
curvature of the field (aberration), 252
cylindrical coordinates, 3

degree of coherence, 206, 207
degree of polarization, 143, 157, 160
density of modes, 322
depth of focus, 293
determinant, 12
dielectric, 43
diffraction grating, 187, 286
diffraction of a Gaussian field profile, 290
diffraction with cylindrical symmetry, 267
diffraction, Fraunhofer, 266
diffraction, Fresnel, 264
diffraction, Fresnel-Kirchhoff, 269
Dirac delta function, 17
dispersion, 43, 169, 180, 181
dispersion relation, 44
dispersion relation in crystals, 125
displacement current, 32
distortion, 252
divergence, 3
divergence theorem, 5
dot product, 2
double interface, 90

eikonal equation, 229, 230, 232
Einstein A and B coefficients, 326
Einstein, Albert, 327
electric field, 26
electric field in a crystal, 135
ellipsometry, 155
elliptical polarization, 143–146
ellipticity, 147, 155
emissivity, 320
energy density, 65, 67, 189, 320
energy exchange between a pulse and medium, 189
energy transport velocity, 190
equipartition principle, 321
Euler's formula, 7
evanescent waves, 82
extraordinary index, 128
extraordinary ray, 121, 130, 139
eye piece, 246

f-number, 294
Fabry, Charles, 98
Fabry-Perot etalon, 98, 101
Fabry-Perot interferometer, 98, 100
Fabry-Perot setup, 100
Fabry-Perot, free spectral range, 102
Fabry-Perot, resolution, 102
far field, 259, 266
Faraday's law, 25, 29, 44
Faraday, Michael, 29
fast axis of a wave plate, 151
Fermat's principle, 229, 233
Fermat, Pierre, 233
finesse, 104
finesse, coefficient of, 95, 99
fluence, 177, 205
focal length, 241, 244
Fourier expansion, 14
Fourier integral theorem, 14, 16
Fourier spectroscopy, 209
Fourier theory, 13
Fourier transform, 16, 175
Fraunhofer approximation, 266
Fraunhofer diffraction with a lens, 277
Fraunhofer, Joseph, 266
free spectral range, Fabry-Perot, 102
frequency, 44
frequency spectrum, 174
Fresnel approximation, 264
Fresnel coefficients, 77, 78
Fresnel zone plate, 311
Fresnel's equation, 125
Fresnel, Augustin, 77
Fresnel-Kirchhoff diffraction, 269
fringe pattern, 305, 306
fringe visibility, 207, 208

fringes, 101
frustrated total internal reflection, 97

Gabor, Dennis, 308
Galileo, 243
Gauss' law, 25, 26
Gauss' law for magnetic fields, 27
Gauss, Friedrich, 27
Gaussian field, diffraction of, 290
Gaussian function, 24
Gaussian laser beam, 292
Gouy shift, 293
gradient, 3
grating, 286
grating pair, 187
Green's theorem, 272
group delay, 170, 178, 183, 184
group delay function, 180
group velocity, 169, 172, 173, 180

half wave plate, 152
Hamilton, William Rowen, 174
Hankel transform, 268
helicity, 147, 155
Helmholtz equation, 262
hologram, generation of, 307
hologram, illumination of, 308
holography, 305
Huygens' principle, 260
Huygens, Christiaan, 130, 259
Huygens, elliptical wavelets in crystals, 138
Huygens-Fresnel, 260
hyperbolic cosine, 7
hyperbolic sine, 7

identity matrix, 12
image formation, 235, 237, 243
imaginary number, 7
imaginary part, 9
index of refraction, 43, 46, 47
instantaneous frequency, 199
instantaneous power spectrum, 193
integrals, 19
interferograms, 305
inverse Fourier transform, 175
inverse matrix, 11
irradiance, 55, 56
isotropic medium, 57, 121

Jeans, James Hopwood, 323
jinc, 269, 283
Jones matrices, 143, 147, 150, 151
Jones vector, 143, 146
Jones vectors, 145
Jones, R. Clark, 145

Kirchhoff, Gustav, 319
Kramers-Kronig relations, 194
Kronecker delta function, 214

Land, Edwin H., 151
Laplacian, 4
laser, 320, 328
laser beam, 292
laser cavity, 248
law of reflection, 75
lens, 240
lens maker's formula, 241
linear algebra, 11
linear medium, 47
linear polarization, 143, 144
Lorentz model, 49, 50
Lorentz, Hendrik, 49
Lorentz-Lorenz formula, 63

magnetic field, 27
magnification, 244
magnitude, 1
matrix multiplication, 11
Maxwell's equations, 25
Maxwell, James Clerk, 32
Michelson interferometer, 203
Michelson, Albert, 205
mirage, 232, 255
Mueller matrix, 160
multilayer coatings, 105
multilayer stacks, 109
multimode, 292

narrowband, 169

negative crystal, 128
Newton, Isaac, 230
nonlinear optics, 41
normal to a surface, 6

object, 237
objective lens, 246
obliquity factor, 264
optic axes of a crystal, 127
optical activity, 164
optical axis, 230, 237
optical path length, 234
optical systems, 246
ordinary, 128
oscillator strength, 52

p-polarized light, 74, 91
paraxial approximation, 230, 237
paraxial rays, 230, 236
paraxial wave equation, 265, 266
Parseval's theorem, 18, 176
partially polarized light, 156
pellicle, 99
permeability, 25
permittivity, 25, 47
phase delay, 180
phase velocity, 172, 180
photometry, 58
photon, 320
Planck blackbody formula, 323
Planck, Max, 325
plane of incidence, 74
plane wave, superposition, 170
plane waves, 43, 45
plasma frequency, 51
Poisson's spot, 262
polarizability, 62
polarization current, 34
polarization effects at an interface, 153
polarization of a material, 34
polarization of light, 143, 144
polarization, partial, 156
polarizer, 143, 147, 150
Polaroid, 147
positive crystal, 128
power spectrum, 175
Poynting vector, 55, 57
Poynting vector in a crystal, 130
Poynting's theorem, 54, 189, 190
Poynting, John Henry, 54
principal axes of a crystal, 123, 126
principal planes, 230, 246
principal value, 195
propagation, wave packet, 178
pulse chirping, 181, 187
pulse stretching, 181

quadratic dispersion, 181
quarter-wave plate, 152, 155

radiometry, 58
radius of curvature, 241
ray, 229, 233
ray diagram, 244
ray tracing, 237, 250
Rayleigh criterion, 283
Rayleigh range, 293
Rayleigh, Lord, 173
rays, reflection/refraction at curved surfaces, 238
real image, 244
real part, 9
rectangular aperture, 265, 267
reflectance, 78
reflection from a curved surface, 238, 239
reflection from a metal, 83
refraction, 73, 75
refraction for a crystal, 129
refractive index, 46
reshaping delay, 185
resolution, 277, 282
resolution, Fabry-Perot, 102
resolution, telescope, 282
resolving power, 104, 289
retarder, 151
right-hand rule, 3
ring cavity, 248
Roemer, Ole, 41

rotation of coordinates, 133

s-polarized light, 74, 91
Savart, Felix, 28
scalar diffraction, 260, 262
scalar Helmholtz equation, 263
Sellmeier equation, 52
senkrecht, 74
signal front, 189
sinc, 267
sine, complex representation, 7
skin depth, 48
slow axis of a wave plate, 151
Snell's law, 75, 129
Snell, Willebrord, 75
spatial coherence, 203, 211, 215
spatial filter, 317
spectrometer, 287
spectrum, 174
spherical aberration, 251
spherical surface, 238, 239
spherical wave, 260, 309
stability of laser cavities, 248
Stefan-Boltzmann law, 320, 328
stochastic phase, 214
Stokes parameters, 159
Stokes vector, 157, 160
Stokes' theorem, 6
Stokes, George Gabriel, 157
Strutt, John William, 173
subluminal, 186
sums, 19
superluminal, 169, 186
surface figure, 306
susceptibility, 47
susceptibility tensor, 122, 132
Sylvester's theorem, 12, 109

table of integrals and sums, 19
Taylor's series, 7
telescope, 246
telescope, resolution of, 282
temporal coherence, 203, 204
testing optical surfaces, 306
thin lens, 241
total internal reflection, 81
transmittance, 78, 79
transmittance through a double interface, 94
tunneling of evanescent waves, 97

uncertainty principle, 177
uniaxial crystal, 127, 128
unit vector, 1
unpolarized light, 143, 156

van Cittert-Zernike theorem, 216
vector calculus, 1
vector multiplication, 2
virtual image, 257
visibility of fringes, 207
voltage, 29

wave equation, 36
wave number, 44
wave packet propagation, 178
wave plates, 143, 151
wavelength, 44
Wien, Wilhelm, 324

Young's two-slit setup, 211
Young, Thomas, 214

zone plate, 311

Physical Constants

Constant	Symbol	Value
Permittivity	ϵ_0	$8.8542 \times 10^{-12}\ \mathrm{C}^2/\mathrm{N}\cdot\mathrm{m}^2$
Permeability	μ_0	$4\pi \times 10^{-7}\ \mathrm{T}\cdot\mathrm{m/A}$ (or $\mathrm{kg}\cdot\mathrm{m}/\mathrm{C}^2$)
Speed of light in vacuum	c	$2.9979 \times 10^{8}\ \mathrm{m/s}$
Charge of an electron	q_e	$1.602 \times 10^{-19}\ \mathrm{C}$
Mass of an electron	m_e	$9.108 \times 10^{-31}\ \mathrm{kg}$
Boltzmann's constant	k_B	$1.380 \times 10^{-23}\ \mathrm{J/K}$
Plancks constant	h	$6.626 \times 10^{-34}\ \mathrm{J}\cdot\mathrm{s}$
	$\hbar$	$1.054 \times 10^{-34}\ \mathrm{J}\cdot\mathrm{s}$
Stefan-Boltzmann constant	σ	$5.670 \times 10^{-8}\ \mathrm{W/m}^2\cdot\mathrm{K}^4$

Integrals and Sums

$$\int_{-\infty}^{\infty} e^{-ax^2+bx+c}\,dx = \sqrt{\frac{\pi}{a}}\,e^{\frac{b^2}{4a}+c} \qquad (\mathrm{Re}\{a\} > 0) \tag{0.55}$$

$$\int_0^{\infty} \frac{e^{iax}}{1+x^2/b^2}\,dx = \frac{\pi\,|b|}{2}\,e^{-|ab|} \qquad (b > 0) \tag{0.56}$$

$$\int_0^{2\pi} e^{\pm ia\cos(\theta-\theta')}\,d\theta = 2\pi J_0(a) \tag{0.57}$$

$$\int_0^{a} J_0(bx)\,x\,dx = \frac{a}{b}J_1(ab) \tag{0.58}$$

$$\int_0^{\infty} e^{-ax^2} J_0(bx)\,x\,dx = \frac{e^{-b^2/4a}}{2a} \tag{0.59}$$

$$\int_0^{\infty} \frac{\sin^2(ax)}{(ax)^2}\,dx = \frac{\pi}{2a} \tag{0.60}$$

$$\int \frac{dy}{\left[y^2+c\right]^{3/2}} = \frac{y}{c\sqrt{y^2+c}} \tag{0.61}$$

$$\int \frac{dx}{x\sqrt{x^2-c}} = -\frac{1}{\sqrt{c}}\sin^{-1}\frac{\sqrt{c}}{|x|} \tag{0.62}$$

$$\int_0^{\pi} \sin(ax)\sin(bx)\,dx = \int_0^{\pi} \cos(ax)\cos(bx)\,dx = \frac{\pi}{2}\delta_{ab} \qquad (a, b \text{ integer}) \tag{0.63}$$

$$\sum_{n=0}^{N} r^n = \frac{1-r^{N+1}}{1-r} \tag{0.64}$$

$$\sum_{n=1}^{N} r^n = \frac{r(1-r^N)}{1-r} \tag{0.65}$$

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r} \qquad (r < 1) \tag{0.66}$$